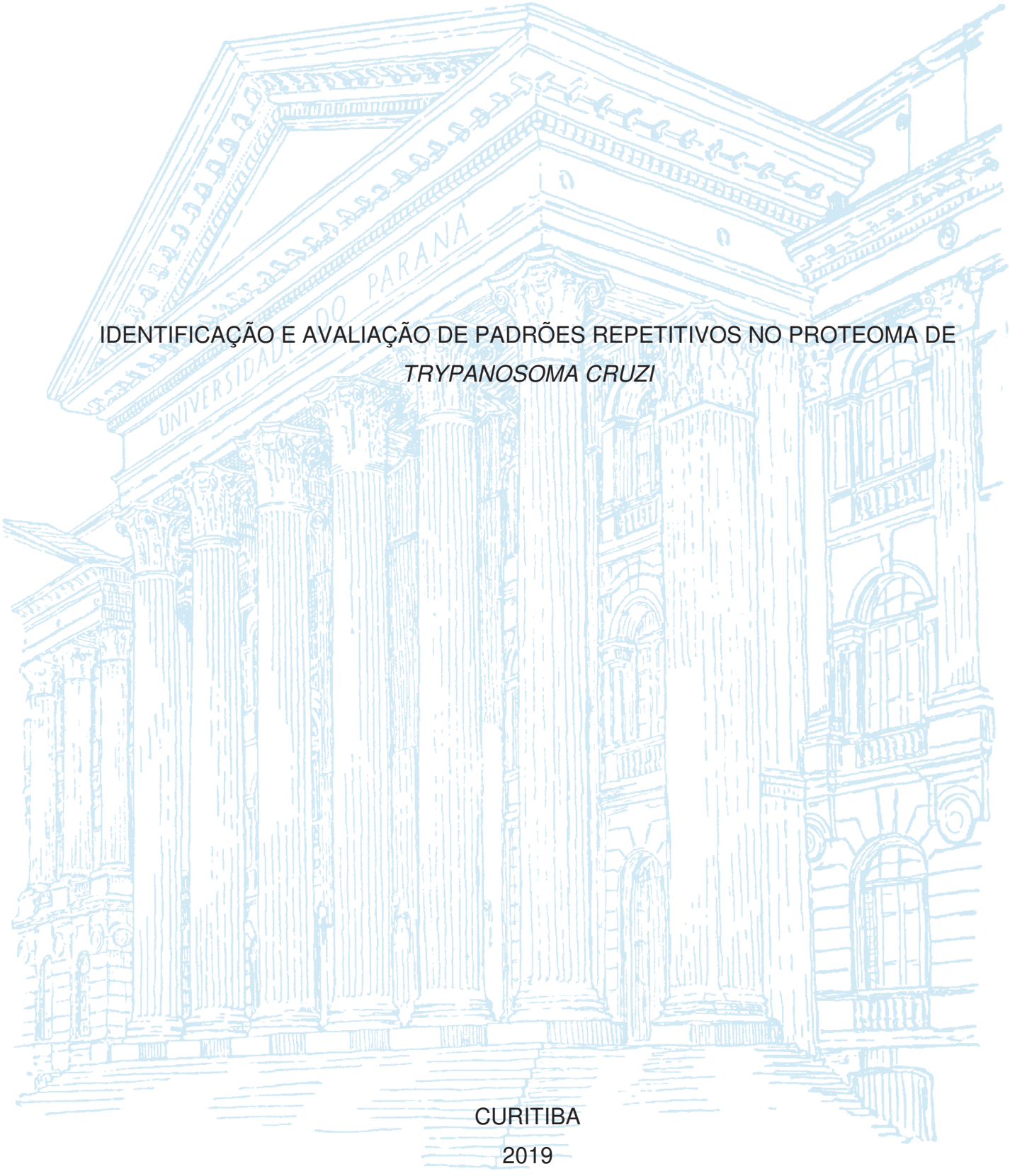
UNIVERSIDADE FEDERAL DO PARANÁ

MARIANE GONÇALVES KULIK



IDENTIFICAÇÃO E AVALIAÇÃO DE PADRÕES REPETITIVOS NO PROTEOMA DE *TRYPANOSOMA CRUZI*

CURITIBA

2019

MARIANE GONÇALVES KULIK

IDENTIFICAÇÃO E AVALIAÇÃO DE PADRÕES REPETITIVOS NO PROTEOMA DE *TRYPANOSOMA CRUZI*

Dissertação apresentada ao curso de Pós-Graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Mestre em Bioinformática.

Orientador: Prof. Dr. Doutor Roberto Tadeu Raittz

Coorientador: Prof. Dr. Wanderson Duarte da Rocha

CURITIBA

2019

Sistema de Bibliotecas/UFPR
Biblioteca de Educação Profissional e Tecnológica

Kulik, Mariane Gonçalves
K96 Identificação e avaliação de padrões repetitivos no proteoma de *Trypanosoma Cruzi* / Mariane Gonçalves Kulik. - Curitiba, 2019.
106 p.: il.

Dissertação (Mestrado) – Universidade Federal do Paraná, Setor de Educação Profissional e Tecnológica, Curso de Pós-Graduação em Bioinformática, 2019.
Orientador: Roberto Tadeu Raittz
Coorientador: Wanderson Duarte da Rocha

1. *Trypanosoma Cruzi*. 2. *Chagas, Doença de*. 3. Inteligência artificial. 4. Bioinformática. I. Raitz, Roberto Tadeu. II. Rocha, Wanderson Duarte da III. Título. IV. Universidade federal do Paraná.

CDD: 616.9363

Angela Pereira Farias Mengatto - CRB9/1002

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA

Pós-Graduação em Bioinformática WWW.BIOINFO.UFPR.BR
E-mail: bioinfo@ufpr.br **Tel: 41 33614906**

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em BIOINFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da dissertação de Mestrado de **MARIANE GONÇALVES KULIK** intitulada: **"Identificação e avaliação de padrões repetitivos no proteoma de *Trypanosoma cruzi*"**, após terem inquirido a aluna e realizado a avaliação do trabalho, são de parecer pela sua Aprovação no rito de defesa.

A outorga do título de mestre está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

Curitiba, 28 de maio de 2019.

Dr. Roberto Tadeu Raittz
Presidente/Programa de Pós-graduação em Bioinformática – UFPR

Dr. Flávio Bortolozzi
Avaliador Externo/Centro Universitário de Maringá-UNICESUMAR

Dr. Fabio de Oliveira Pedrosa
Avaliador Interno/Programa de Pós-graduação em Ciências-Bioquímica/
Programa de Pós-graduação em Bioinformática-UFPR

Dr. Wanderson Duarte da Rocha
Avaliador Interno/Programa de Pós-graduação em Ciências-Bioquímica/
Programa de Pós-graduação em Bioinformática-UFPR

Dedico esta dissertação ao meu marido Fausto Allan Kulik
e a minha grande família (original e adquirida).
Amo vocês!

AGRADECIMENTOS

Ao Professor Dr. Roberto T. Raittz pela dedicação, pela constante ajuda, pelo enorme aprendizado, por toda confiança que depositou em mim e pela liberdade conferida para realização deste projeto.

Ao Professor Dr. Wanderson D. DaRocha pela paciência, grande aprendizado e todo o esforço em buscar uma língua comum entre nossos 2 mundos para me passar conhecimento.

Ao meu amado esposo Fausto Allan Kulik, pela cumplicidade, compreensão, enorme paciência e pelo exercício diário para tornar possível a realização deste trabalho.

Aos meus pais Mauro José Gonçalves e Eloá Maria Gonçalves, sou eternamente grata por terem me tornado o que sou, pelos cuidados e amor incondicional; e aos meus irmãos Fabiane Gonçalves, pelas risadas constantes, Eduardo José Gonçalves, por toda a cumplicidade e pelos 3 anjinhos que trouxe para alegrar nossas vidas, e Viviane Gonçalves (*in memorian*), que apesar de ter permanecido tão pouco em nossas vidas, sempre será lembrada e amada.

Aos demais amigos e parentes, originais ou adquiridos (em especial a minha sogra e "mãe de coração" Lorena Maria Alberti Kulik), por todo o carinho, companheirismo e momentos felizes compartilhados.

As amigas que fiz no laboratório de Bioinformática da UFPR, Sheyla Trefflich e Hellen Cristine Machado, pelos momentos bons (e de desespero também) compartilhados e pela parceria em todo o período do mestrado.

A todos os professores do programa de Bioinformática da UFPR, em especial ao professor Dr. Dieval Guizelini e à secretária Suzana Gobetti, por todo o suporte e conhecimento repassados, que garantiram o sucesso deste trabalho.

Aos colegas do laboratório de Bioinfo pela prazerosa convivência, em especial pela hora do café, e toda ajuda oferecida.

A CAPES, pelo apoio financeiro dado.

A percepção do desconhecido é a mais fascinante das experiências.
O homem que não tem os olhos abertos para o misterioso passará pela vida
sem ver nada.

ALBERT EINSTEIN

RESUMO

Regiões repetidas são primordiais para a sobrevivência de *Trypanossoma cruzi*, pois tem sido atribuído a elas um papel importante no processo de evasão do sistema imune de hospedeiros mamíferos. A compreensão de muitas das funções que estas características exercem e/ou mecanismos ainda nos escapa, fazendo com que a patologia causada por este parasito, a Doença de Chagas, ainda não tenha cura definitiva. Alem disso, o diagnóstico desta doenca ainda é limitante. Neste estudo, nós aplicamos novas técnicas de bioinformática para anotação e análise exploratória de *tandem repeats* (TRs). Nós também realizamos análises de preferência de códons no proteoma completo e aplicamos valores de cobertura de transcriptoma para avaliarmos possíveis diferenças de seleção de códons para diferentes etapas do ciclo de vida. Verificamos que alguns aminoácidos apresentaram divergências, porém a grande maioria apresenta as preferências do genoma, enquanto que entre as etapas do ciclo de vida os padrões são sempre os mesmos. Ao compararmos as preferências globais com as regiões de TRs, verificamos que nas proteínas transmembrana, elas apresentam características distintas que podem indicar um meio de suprimir a expressão destes genes. Aprofundando nossas análises de TRs, nós realizamos a anotação de epitopos de células B nessas regiões e aplicamos dados de transcritoma buscando os melhores candidatos para novos alvos de teste de diagnóstico. Além de alguns dos antígenos já conhecidos, fomos capazes de identificar outros candidatos promissores a testes experimentais. Ao final do processo nós aplicamos as lições aprendidas com identificação de *Tandem Repeats* na geração de um modelo capaz de classificar sequências com e sem TRs, atingindo acurácia de 80%. O modelo desenvolvido aqui permitirá identificar TRs conservados em outros organismos patogênicos, bons alvos para anotação de epitopos de célula B para testes diagnósticos.

Palavras-chave: *T. cruzi*. *Tandem repeat*. codon. Bioinformática. Inteligência Artificial.

ABSTRACT

Repeated regions are crucial to *Trypanosoma cruzi* survival, since it has been assigned to them an important role in the evasion of the mammalian host immune system. Many of their functions and/or mechanisms remain unknown, and the definitive cure for the desease caused by it, Chagas Desease, still deceives us. In addition, its diagnosis is still limiting. Here, we applied new bioinformatics techniques to annotate and perform exploratory analysis of Tandem Repeats (TRs). We also performed codon preference analysis on the complete proteome and applied transcriptome coverage values to assess differences in codon selection at different stages of the life cycle. We found that some of the amino-acids presented divergency, but most of them share the preferences of the genome, while between the stages of the life cycle the patterns are always the same. When the preferences of the TR regions were compared to the global ones, we found that, for transmembrane proteins, these preferences presented some distinct characteristics, which may suggest a way to suppress these genes. We then annotated B-cell epitopes in these TR regions and applied transcriptome data on them, looking for better targets for diagnostic tests. In addition to some well-known antigens, we were able to find other promising candidates to future experimental testing. As the last task in this process, we applied the lessons learned to find Tandem Repeats in an AI model. It is able to classify sequences with and without TRs, with an accuracy of 80%. This model will allow the identification of sequences with conserved TRs in other pathogenic organisms, which will be good targets for B-cell epitope tools and future diagnostic tests.

Keywords: *T cruzi*. *Tandem repeat*. Codon. Bioinformatics. Artificial Inteligence.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS OU SIGLAS

AA - Aminoácido
AI - Artificial Intelligence
ANN - Redes neurais artificiais
BCR - B cell receptors
CAI - Codon Adaptation Index
CDS - Coding sequence
DNA - Deoxyribonucleic acid
ELISA - Enzyme Linked Immunosorbent Assay
GO - Gene Ontology
IFA - Indirect immunofluorescence assay
IHA - Indirect haemagglutination assay
MASP - Mucin-associated surface protein
MHC - Major histocompatibility complex
ML - Machine Learning
MLP - Multi-layer Perceptron
MR - Módulo da repetição
NT - Nucleotídeo
ORF - Open read frame
PCR - Polymerase chain reaction
PPR - Pattern recognition receptors
RBF - Radial Basis Function
RF - Random Forests
RNA - Ribonucleic acid
RSCU - Relative Synonymous Codon Usage
RPKM - Reads Per Kilobase Million
SVM - Support Machine Vectors
TLR - Toll-like receptors
TR - Tandem Repeat
UTD - Unidades de tipificação discreta
WHO - World Health Organization

SUMÁRIO

1 INTRODUÇÃO

A doença de Chagas é uma doença tropical negligenciada que acomete 7 milhões de pessoas, causando em torno de 10 mil mortes anuais. Trata-se de uma enfermidade endêmica na América Latina e México, porém, devido a globalização os casos tem ocorrido em todo mundo nos últimos anos, sendo detectada na América do Norte, Europa, Ásia e Oceania.

Descoberta por Carlos Chagas em 1909 no Brasil, a doença e seu agente etiológico, *Trypanosoma cruzi*, vêm sendo alvos de diversas pesquisas desde então. No entanto, sua complexidade genotípica e fenotípica, e de sua grande capacidade de evasão do sistema imunológico de hospedeiros mamíferos ainda apresentam grandes desafios a nossa compreensão. Atualmente sua cura só é possível nos estágios iniciais de infecção e sua detecção é difícil, por vezes escapando às técnicas disponíveis.

A análise genética do *T. cruzi* é extremamente complexa devido ao excesso de regiões repetidas, conhecidas como *Tandem Repeats* (TR). Além de uma das principais causas de problemas no fechamento do seu genoma em 2005, estas regiões também dificultam experimentos *in vitro* devido suas características bioquímicas. Não obstante, sua compreensão é peça chave para entendermos mais profundamente os mecanismos e características do parasito.

Um dos diversos papéis dos TR dos quais já temos conhecimento há décadas é o de apresentar características imunogênicas, causando respostas desfocadas do sistema imune do hospedeiro. Estudos exploratórios de TRs com foco em proteínas de superfície, que naturalmente estão envolvidas no processo de adesão celular, já foram realizados, porém ainda temos carência por uma análise mais abrangente do seu potencial antigênico em todo o proteoma.

As técnicas de sorologia evoluíram muito nas últimas décadas, com fortes investimentos em kits comerciais, porém a pesquisa por novos alvos está estagnada desde o final dos anos 2000 e uma revisão dos possíveis alvos, bem como o desenvolvimento de meios mais ágeis de detecção, podem contribuir muito para conhecermos um pouco melhor este organismo complexo e intrigante.

1.1 OBJETIVOS

1.1.1 Objetivo geral

Efetuar uma análise exploratória utilizando ferramentas de bioinformática para anotação de *Tandem Repeats*, verificando suas características gerais e potenciais como epitopos de célula B, ou seja, bons candidatos para desenvolvimento de testes sorológicos.

1.1.2 Objetivos específicos

- Identificar módulos repetitivo (MR) e área de *tandem repeat* (TR) no proteoma de *T. cruzi* CL Brener Emeraldo-like;
- Validar ferramentas de anotação e técnicas de filtragem capazes de caracterizar *Tandem Repeats* com diversidade moderada;
- Avaliar composição geral de aminoácidos e de códons das regiões de *Tandem Repeats;*
- Aplicar informações de cobertura na determinação de códons preferenciais considerando-se os dados transcriptômicos de diferentes etapas do ciclo de vida;
- Predizer os TRs que apresentam prováveis epitopos de célula B com *scores* elevados;
- Desenvolver modelo de inteligência artificial com base nas lições aprendidas em todo o processo executado que permita investigação de Tandem Repeats e simplifique o processo de detecção de epitopos de célula B em outros patógenos eucariotos.

2 REVISÃO DE LITERATURA

2.1 TRYPANOSOMA CRUZI

O parasito Trypanosoma cruzi é o agente da doença de Chagas ou Tripanossomíase Americana, uma doença tropical que afeta milhares de pessoas de forma endêmica na América Latina e México (Moncayo; Silveira (2009)). Sua transmissão ocorre através da picada de insetos contaminados da subfamília dos Triatomíneos em humanos e animais do seu meio doméstico, como cães e gatos, e animais silvestres, como roedores e marsupiais. Outros meios de infecção são por transfusão de sangue, transplante de órgãos, acidentes laboratoriais, transmissão vertical de mãe para filho e sexual (RASSI *et al.*, 2010). Casos recentes de contaminação por ingestão de alimentos e líquidos, uma das técnicas epidemiológicas mais ancestrais de transmissão para novos hospedeiros (GÜRTLER; CARDINAL, 2015), pode estar associada com o aumento na virulência da doença (SEGOVIA *et al.*, 2013).

2.1.1 Um ciclo de vida complexo

T. cruzi são protozoários flagelados pertencentes à classe Kinetoplastida, família *Trypanosomatidae*. São eucariotos unicelulares compostos de um núcleo diferenciado e contém no citoplasma uma estrutura denominada cinetoplasto, que contém o DNA mitocondrial e é anexa à mitocôndria (SOUZA, 2002).

Tratam-se de parasitos digenéticos que apresentam quatro formas morfológicas durante seu ciclo evolutivo. Diferenciam-se em epimastigotas no tubo digestivo do vetor triatomíneo, onde apresentam forma de fuso e núcleo semi-central anterior ao cinetoplasto. Neste estágio efetuam a divisão por fissão binária até que atinjam a porção final do trato digestivo do hospedeiro, transformando-se em tripomastigotas metacíclicos. Nesta fase possuem forma alongada e cinetoplasto posicionado entre o núcleo e posição posterior do organismo. Sua transmissão inicia-se com as fezes do vetor invertebrado infectado depositadas no tecido epitelial lesionado ou mucosas do hospedeiro mamífero. Seu flagelo tem papel essencial neste processo, bem como para locomoção pós-infecção através de tecidos e internalização por diversos tipos celulares do hospedeiro (SOUZA, 2002).

Durante a infecção/invasão celular, a forma metacíclica sofre uma cascata de sinalizações geradas pelo processo de interação parasito-hospedeiro e diferencia-se em amastigota, escapando para o citoplasma. Neste estágio ele apresenta formato arredondado e achatado, com cinetoplasto visível e pequeno flagelo. Em meio citoplasmático, esta forma reprodutiva inicia o processo de multiplicação por fissão binária e após diversas rodadas de replicação diferencia-se em tripomastigota sanguíneo (não replicativo). Esta é altamente móvel, apresentando uma forma alongada, porém menor e mais robusta do que a apresentada pela forma metacíclica, com flagelo proeminente (SOUZA, 2002).

Os tripomastigotas sanguíneos causam a ruptura da membrana celular do hospedeiro, escapando para a corrente sanguínea atingindo novos tecidos. Novos triatomíneos são infectados através do repasto de hospedeiros mamíferos e o ciclo se reinicia com a diferenciação em epimastigota na região intestinal do hospedeiro invertebrado (SOUZA, 2002).

2.1.2 A Doença de Chagas

Apesar das medidas de conscientização adotadas para redução dos insetos vetores em países onde a doença é endêmica, temos registros de que até 2009 o número de mortes anuais chegava a 6 mil casos (MARTINS-MELO *et al.*, 2012). Segundo A Organização Mundial de Saúde (WHO) em uma estimativa mais atualizada, indica 6 a 7 milhões de pessoas infectadas e em torno de 10 mil mortes anuais. Casos de infectados pela doença foram reportados na América do Norte, Europa, Ásia e Oceania, devido ao aumento das imigrações ocasionadas pela globalização (GASCON *et al.*, 2010; JACKSON *et al.*, 2014). De acordo com (CAPUANI *et al.*, 2017), temos um cenário ainda mais alarmante, onde pacientes não diagnosticados devido ao estágio da doença denominado estado crônico indeterminado podem ter a doação de sangue autorizada, o que geraria números ainda maiores de infectados.

A doença de Chagas apresenta as fases aguda e crônica. A primeira tem duração de 30 a 90 dias, e tem como principal característica um grande número de tripomastigotas sanguíneas circulantes, apresentando sintomas variáveis, principalmente febre, dores de cabeça e fraqueza e, em casos mais raros, nódulos na pele (chagomas) e edema prolongado na região da picada (sinal de Romaña).

Em menos de 1% dos casos há risco de morte nesta fase, em decorrência de meningo-encefalite ou miocardite. A cura parasitológica ocorre em torno de 70% dos indivíduos diagnosticados logo após a infecção (revisado por BERN, 2015).

Os pacientes onde a doença persiste evoluem para a fase crônica, onde os níveis de parasitemia ficam reduzidos. De 60 a 70% progridem para um estado de aparente equilíbrio parasito-hospedeiro, sem manifestações clínicas, conhecido como crônico indeterminado (assintomático). Com o tempo, estes pacientes podem desenvolver uma forma sintomática da doença, que se caracteriza pelo baixo nível de parasitos e quantidade elevada de anticorpos circulantes (revisado por DUTRA *et al.*, 2014). 30-40% do total de infectados evoluem para a forma crônica cardíaca, manifestando miocardite crônica ou hipertrofia do coração, e 10% para a digestiva, apresentando hipertrofia no esôfago e cólon intestinal. Há também casos de co-ocorrência das duas formas, enquanto que demais infectados podem conviver com a doença até o fim da vida (RASSI et al., 2010).

2.1.3 Genoma e regiões repetidas

A capacidade de adaptação dos tripanosomatídeos foi sem dúvida um fator preponderante à sobrevivência destes patógenos por milhares de anos. Com base na classificação mais atual dispomos de 6 DTUs reconhecidas (UTDs - Unidades de tipificação discreta) e uma sétima proposta, conhecida como TcBat (revisado por ZINGALES, 2018).

O primeiro processo de montagem do genoma do *T. cruzi* foi publicado por (El-Sayed (2005)). Este utilizou a cepa híbrida CLBrener, ou seja seu genoma pode ser organizado em dois haplótipos, o haplótipos “Esmeraldo-like”, devido a sua semelhança com o genomas da cepa Esmeraldo (DTU II), e “non Esmeraldo-like”, ou seja sequências mais divergentes do haplótipo Esmeraldo-like. A montagem do genoma foi fortemente prejudicada por esta condição híbrida, além de sua alta composição de regiões repetidas, chegando a 50%, e múltiplas cópias de proteínas de superfície, sendo as maiores as de mucinas, proteínas de superfície associadas a mucinas (MASP), trans-sialidases e glicoproteínas de superfície gp63 (EL-SAYED, 2005).

Novas tentativas de sequenciamento e montagem de genoma com outras cepas foram realizados nos últimos anos (BAPTISTA *et al.*, 2018; BERNÁ *et al.*,

2018; FRANZÉN et al., 2011), porém tradicionalmente, o haplótipo “Esmeraldo-like” ainda vem sendo muito utilizado nas análises *in vivo*, *in vitro* e *in silico* e a questão e regiões repetidas ainda permanece um desafio, pois dificulta o processo de fechamento do genoma e consequentemente suas análises.

2.1.4 Particularidades na regulação gênica

Nos eucariotos em geral, o processo de diferenciação celular é controlado em múltiplos níveis, por exemplo, ao nível transcricional pela ação de fatores de iniciação de transcrição. Já em tripanosomatídeos são considerados ausentes os mecanismos de controle da expressão via regulação da RNA polimerase II, apesar de já ter sido descrita a participação das variantes de histonas específicas na transcrição de unidades policistrônicas (revisado por CLAYTON, 2016). Íntrons são extremamente incomuns, tendo sido detectados apenas 2 genes que sofrem o processo de cis-splicing, processo realizado somente em regiões onde estes são presentes (MAIR et al., 2000).

A ausência de mecanismos de regulação na etapa de transcrição é compensada em etapas seguintes do processo. Os RNAs destinados a codificação de proteínas sofrem um processamento na etapa pós-transcricional de “trans-splicing”. Nela as todas as unidades de RNA mensageiro (mRNA), são geradas à partir de transcritos independentes conhecidos como pré-mRNAs, que são clivados e recebem uma região adicional de ~39 NTs conhecida como “spliced leader” (SL) na Cap 5’ (VAN DER PLOEG, 1986). Posteriormente, na região 3’ ocorre o processo de poliadenilação, requerido para maturação do mRNA e exportação ao citoplasma (ULLU et al., 1993).

Análises in silico realizadas por Horn (2008) apontaram evidências de seleção traducional, onde códons mais frequentes correspondentes a tRNAs mais abundantes teriam papel importante no processo de repressão e ativação da expressão gênica. Este tipo de pressão seletiva também foi identificado em outros organismos anteriormente (AKASHI; EYRE-WALKER, 1998;DOS REIS; WERNISCH, 2009), porém os baixos níveis de regulação em outras etapas do processo, como o nível transcricional, sugere que o uso preferencial de codons possa ser um mecanismo importante no controle traducional em tripanossomatídeos.

2.1.5 A complexa interação patógeno-hospedeiromamífero

As formas tripomastigotas metacíclicas são capazes de invadir uma série de célula nucleadas do hospedeiro mamífero. Para isto, *T. cruzi* dispõem de um arsenal de proteínas polimórficas de superfície ancoradas com glicosilfosfatidilinositol (GPI), como as mucinas, MASPs e trans-sialidases que viabilizam este processo. Os componentes clássicos do sistema imune inato, como macrófagos, células dendríticas e as *Natural Killers* (NK) tem papel crucial nesta fase inicial, no entanto a infecção é capaz de persistir (EPTING *et al*., 2010). Em estudo experimental realizado por (PADILLA *et al*., 2009) em camundongos, verificou-se que nas primeiras 24hs da infecção uma quantidade pequena de parasitos chega aos locais de drenagem dos nódulos linfáticos, indicando uma falha no sequestro nas regiões infectadas. Este pode ser um indicativo do mecanismo de evasão do parasito, que infecta células não-profissionais, que usualmente não são utilizadas no processo de sinalização do sistema imune, atrasando as etapas mais avançadas de especialização de anticorpos (PADILLA *et al*., 2009).

O parasito circulante é internalizado através da invaginação da membrana plasmática da célula hospedeira em uma estrutura denominada fagolisossomo. Esta possui uma composição extremamente ácida que tem como objetivo digerir agentes patogênicos no processo de fagocitose. Este ambiente oxidativo atua como sinalizador de diferenciação para a forma amastigota, ocasionando a ruptura da membrana do fagolisossomo em torno de 24hs após infecção, garantindo o baixo nível de parasitemia nas primeiras horas de infecção (PIACENZA *et al*., 2013).

Em *T. cruzi*, moléculas de superfície como mucinas e glicosol fosfolipídeos (GPL) tem grande capacidade estimuladora de TLRs, que são receptores anexados à superfície das células do sistema imune inato (PRRs - *pattern recognition receptors*). Os PPRs do tipo Toll (TLR) são os mais estudados e são os responsáveis pela produção de citocinas e quimiocinas pró-inflamatórias, que atuam como recrutadoras de células fagocíticas para o local da infecção, além de auxiliar no processo de moldagem da resposta do sistema imunológico adaptativo (revisado por TAKEUCHI; AKIRA, 2010). A capacidade de evasão do parasito nas primeiras 24 hs de infecção ocasiona uma baixa parasitemia que atrasa o processo de maturação da reação adaptativa, por não estimular os TLRs, responsáveis pela sinalização necessária à ativação dos linfócitos B (produtor de anticorpos), linfócitos

T CD4+ e T CD8+, principais linhas de defesa especializada que atua quando a infecção resiste a resposta do sistema imunológico inato (DE AVALOS *et al*., 2001).

Após diversas rodadas de infecção e proliferação, o sistema imune é capaz de desenvolver uma resposta robusta, porém insuficiente para eliminação da infecção. Isso pode estar relacionado ao grande repertório de proteínas polimórficas e imunogênicas de superfície co-expressas pelo parasito (DE AVALOS *et al.*, 2001).

Em *Plasmodium falciparum*, o processo de variação antigênica ocorre com a expressão de variantes idênticas na maioria das células dos parasitas da população, enquanto alguns expressam variantes diferentes. O sistema imune adaptativo gera anticorpos especializados na variante mais presente e falha na eliminação completa da parasitemia (GUIZETTI; SCHERF, 2013). Atualmente não dispomos de evidência de que *T.cruzi* adote esse processo, porém acredita-se que o grande repertório de mucinas, MASPs e trans-sialidases ocasione uma resposta do sistema imune com ataques espúrios e não neutralizantes, conhecido como “Cortina de fumaça” (PITCOVSKY *et al.,*2002). Outro fator que ainda exige verificações in vivo, mas pode contribuir com o processo de geração de anticorpos de baixa afinidade, é a presença de mitógenos de célula B derivados de proteínas do parasito e estruturas imunogênicas compostas por longos tandem repeats, podendo ambas causarem a ativação policlonal de células B inespecíficas no hospedeiro (REINA-SAN-MARTIN *et al*., 2000).

Mendes *et al.* (2013), efetuou um estudo da presença de regiões de *Tandem Repeats* em protozoários patogênicos intra e extracelulares e em protistas de vida livre e verificou uma maior concentração TRs e nos organismos patogênicos, em especial em proteínas de superfície, um forte indício de que regiões repetidas desempenham um papel importante no processo de evasão do sistema imune do hospedeiro. Descreveremos em detalhes os processos de bioinformática utilizados para detecção de TRs na seção 2.2.1.

2.1.6 Métodos de diagnóstico existentes

Os métodos de diagnóstico da Doença de Chagas variam de acordo com a fase clínica da doença. Durante a fase aguda e para casos de transmissão congênita, o diagnóstico por microscopia, onde tripomastigotas circulantes podem ser visualizados na corrente sanguínea, ainda é o método mais economicamente

viável. Caso este apresente resultado negativo, testes de concentração, como microhematócrito ou Strout que apresentam sensibilidade de 80 a 90%, devem ser realizados (GOMES et al., 2009; DIAS *et al.*, 2016).

Já na fase crônica indeterminada, que inicia-se imediatamente após a aguda, o nível de parasitemia é reduzido, o que compromete o diagnóstico parasitológico (DE AVALOS *et al.,* 2001). Os métodos indiretos, como xenodiagnóstico e hemocultura, são caros, difíceis, apresentam sensibilidade baixa sensibilidade (20 a 50%) e não são efetivos para alguns pacientes.

A abordagem sorológica passou então a ser a mais aceita e indicada atualmente para este estágio da doença. Ela baseia-se na concentração de anticorpos IgG que ligam-se especificamente a antígenos de *T. cruzi* (DIAS et al., 2016). Segundo revisado por Balouz *et al*., (2017), dispomos de uma variedades de testes sorológicos e podemos destacar:

- Teste de hemaglutinação indireta (IHA - *Indirect haemagglutination assay*): Alta sensibilidade, baseada na aglutinação de hemácias sensibilizadas com antígeno apresentado em solução;
- Teste de imunofluorescência indireta (IFA - *Indirect immunofluorescence assay*): Alta sensibilidade, baseada na reação da forma epimastigota de *T. cruzi* com anticorpos presentes no soro;
- ELISA (*Enzyme Linked Immunosorbent Assay*) com antígeno total: Alta sensibilidade e baixa especificidade, baseada na interação anticorpo-antígeno por meio de uma enzima, utiliza antígenos isolados e tem como ponto negativo a reação cruzada com antígenos de outros tripanossomatídeos;
- ELISA com antígenos recombinantes: Alta especificidade, porém baixa sensibilidade, utiliza frações antigênicas recombinantes ou peptídeos quiméricos contendo regiões repetidas reativas nas interações enzimáticas, que tem como vantagem reação cruzada reduzida. Suas pesquisas costumam utilizar um repertório de diferentes combinações de códons com o objetivo de aumentar a síntese de peptídeos através dos códons ótimos. A lista de recombinantes com suas informações de composição, quando disponível, foi compilada no ANEXO 1 (SANTOS *et al.*, 2016).

Ainda segundo Balouz *et al*. (2017), os ensaios com ELISA de antígeno total antecederam uma onda de iniciativas que, ao final dos anos 1980, dedicaram-se a validação do repertório antigênico baseadas em genes recombinantes. *Kits*

comerciais, contendo amostras únicas ou quiméricas com peptídeos recombinantes foram desenvolvidos e permanecem até hoje sendo os mais utilizados. Novos *kits* que combinam técnicas e alvos comerciais foram lançados porém, apesar de promissores, não trazem abordagens inovadoras para esta área de pesquisa.

Apesar da sua diversidade, nenhum dos testes sorológicos é considerado referência, pois todos apresentam aspectos negativos com relação a reprodutividade, confiabilidade, sensibilidade ou especificidade. Por essa razão, o II Consenso Brasileiro em Doença de Chagas e WHO recomendam testes sorológicos aos pares: um com alta sensibilidade e um com alta especificidade, sendo necessários re-testes caso haja discordância nos resultados (DIAS *et al*., 2016; BALOUZ et al., 2017).

Com o objetivo de combater os problemas de especificidade e sensibilidade inerentes aos testes sorológicos, no final dos anos 80, com a introdução da nova técnica de amplificação por PCR, métodos moleculares foram desenvolvidos. Estes tinham como alvos inicialmente o DNA mitocondrial presente no cinetoplasto (SIMPSON, 1986) e mais recentemente aptâmeros de RNA estáveis (NAGARKATTI*et al.*, 2012). Ramírez *et al.* (2015) efetuaram testes para verificação dos métodos moleculares disponíveis, obtendo sensitividade clínica de ~80%, que não supera a dos testes sorológicos. Estes métodos, além de mais complexos, são afetados pela flutuação de parasitemia o que faz com que os testes sorológicos ainda sejam os mais indicados.

Estudos atuais voltados para detecção de novos alvos para testes sorológicos em *T. cruzi* baseiam-se na síntese de peptídeos recombinantes quiméricos, como o efetuado por Hernández *et al.* (2010). Esta técnica permite combinação de diferentes códons em sua composição, com o objetivo de otimizar o processo de geração de proteínas mais adaptadas a disponibilidade de tRNAs (MAURO; CHAPPELL, 2014). As técnicas de análises de preferências de códon serão descritas em mais detalhes na seção 2.2.1, enquanto as de detecção in silico de epitopos de célula B, alvos dos testes sorológicos, serão explicadas na seção 2.2.3.

2.2 ABORDAGENS DE BIOINFORMÁTICA

2.2.1 Predição de preferência de códons

Diversas métricas baseadas em análises matemáticas foram desenvolvidas com o objetivo de predizer a preferência de códons dos organismos. Elas foram inicialmente desenvolvidas com o intuito de auxiliar na verificação de erros no sequenciamento de DNA e checagem de ORFs geradas ao acaso, algumas datando do início da década de 1980 (revisado por ROTH *et al.,* 2012).

Índices de utilização de códon são geralmente gerados como um número único que mede o quanto um gene está mais fortemente adaptado aos códons preferenciais, gerados à partir de um conjunto de referência. Entre as métricas existentes, a mais comumente utilizada é o CAI (*Codon Adaptation Index*), que utiliza como conjunto de referência a distribuição de códons mais comum aos genes diferencialmente expressos em transcriptoma ou proteoma, conhecida como RSCU (*Relative Synonymous Codon Usage*). O RSCU atribui um peso para cada um dos códons de acordo com sua relevância no grupo de referência, utilizando a premissa de que os genes mais expressos o são devido a conterem o conjunto mais adaptado de códons. O CAI aplica esses pesos aos códons, gene a gene, e calcula uma média geométrica, gerando um valor único, que indica o quanto o gene está adaptado aos códons preferenciais (revisado por ROTH et al., 2012). Apesar de sua correlação indireta com níveis de expressão através do conjunto de referência de RSCU, a distribuição efetiva de códons é calculada sobre os dados do genoma e nem sempre é capaz de refletir os níveis de expressão apresentados pelo proteoma.

Jeacock *et al.* (2018) buscaram calibrar o cálculo de CAI com dados de transcriptomas e proteomas como referência para *T. brucei*. A abordagem inversa foi testada por Piovesan *et al.* (2013), distribuindo dados de transcriptoma sobre a contagem de dados do genoma para predição de códons ótimos. Ambas abordagens partem do pressuposto de que os níveis de transcriptoma podem oferecer informações mais reais de preferência de códons do que a contagem à partir do genoma.

As distribuições de códon utilizadas atualmente como referência para *T. cruzi* foram geradas pelo Instituto de Pesquisa em DNA Kazusa, do Japão, à partir de 289 CDS's. Em estudo comparativo realizado sobre dados de tripanossomatídeos, Horn

(2008) propôs uma tabela de distribuição de códons baseada em proteínas de superfície. Não temos conhecimento de estudos deste tipo efetuados sobre todo o conteúdo do proteoma de CLBrener “Esmeraldo-*like*”.

2.2.2 Ferramentas para detecção de TRs

Regiões repetitivas em genomas são classificadas de acordo com suas características e razões pelas quais acredita-se terem sido formadas. Tandem repeats são porções da sequência que apresentam repetições contíguas e podem ser encontradas na fita de DNA e em regiões codificadoras de proteínas. Acredita-se que estas tiveram origem na replicação e recombinação interna de um mesmo gene e tem relação com propriedades estruturais e funcionais da proteína (ANDRADE et al., 2001). Distinguem-se de repetições relacionadas a elementos transponíveis, que tem como principal característica repetições dispersas no genoma, resultados de “saltos” de porções da sequência intra e extra gene, com importante papel evolutivo em eucariotos (LERAT, 2010).

Diversas ferramentas computacionais foram desenvolvidas no decorrer dos últimos 20 anos para análises de regiões de repetição em tandem. O foco inicial foi o de detecção e anotação, com algoritmos baseados em genoma e proteoma. Paralelamente, algoritmos voltados à detecção de TRs com informações da sua estrutura 3D foram desenvolvidos, bem como repositórios dedicados ao seu registro. Além das técnicas de definição do repeat, estas ferramentas contam com uma gama de métricas desenvolvidas com o intuito de avaliar o grau da diversidade interna sem perda de identidade (PELLEGRINI, 2015). Ainda segundo Pellegrini (2015), as ferramentas baseadas em sequência disponíveis dividem-se em 2 grupos: Detecção de k-mers seguida por extensão e programação dinâmica e auto-alinhamento.

O algoritmos baseados em detecção e extensão de k-mers são a alternativa com execução mais rápida e capacidade de anotação de TRs menos diversos. Nesta categoria destacam-se as ferramentas XSTREAM (NEWMAN; COOPER, 2007) e T-REKS (JORDA; KAJAVA, 2009) para NTs e AAs; e TRF (BENSON, 1999) exclusivamente para NTs (PELLEGRINI, 2015; SCHAPER*et al.*, 2012). Apesar de diferenças em suas abordagens, todas apresentam a estratégia de definir um ou vários tamanhos de “palavra”, inferior ao tamanho da sequência (k-mer), e deslizar por toda a sua extensão, buscando por regiões com baixa diversidade. Os k-mers

menos diversos são isolados e então diversas tentativas de extensão são efetuadas de acordo com os limites estabelecidos pelas métricas de avaliação. A ferramenta T-REKS também utiliza a informação de tabelas de substituição de AAs, o que lhe confere capacidade de detecção de TRs mais diversos (JORDA; KAJAVA, 2009).

Já os algoritmos baseados em programação dinâmica e auto-alinhamento tendem a ser mais lentos devido a abordagem adotada e os TRs anotados são propensos a maior divergência (PELLEGRINI, 2015; SCHAPER*et al.*, 2012). Aqui destacamos as ferramentas TRUST (SZKLARCZYK; HERINGA, 2004), RADAR (HEGER; HOLM, 2000) e HHRepID (BIEGERT; SÖDING, 2008) para detecção em AAs; e TRed (SOKOL*et al.*, 2007) e STAR (DELGRANGE; RIVALS, 2004), exclusivos para DNA. Em todos os casos, o cerne do algoritmo consiste em efetuar múltiplos alinhamentos sub-ótimos utilizando técnicas convencionais de programação dinâmica, dispondo de tabelas de substituição de AAs quando cabível, e posterior avaliação de qualidade à partir de auto-alinhamento, com métricas internas para avaliação do grau de degeneração. TRUST e HHRepID apresentam como característica não só a detecção de TRs, mas também a de repetições intercaladas, que tem como propriedade sua presença espaçada dentro do mesmo gene, apresentando módulos não contíguas (PELLEGRINI, 2015; SCHAPER*et al.*, 2012).

Mais recentemente a ferramenta PROGERF foi desenvolvida, com abordagem ligeiramente diferente das demais (LOPES *et al.*, 2015). Ela utiliza o recurso de detecção utilizando janela deslizante e simula degenerações dentro dessas regiões. Ainda não teve avaliação comparativa com outras ferramentas efetuada além das indicadas no próprio artigo.

Com exceção das ferramentas TRF e T-REKS, as demais ferramentas eliminam internamente sobreposições detectadas, fornecendo a melhor opção como resultado final de acordo com seus critérios de filtragem interna. TRF e T-REKS deixam ao encargo do utilizador a decisão de qual o TR mais adequado ao estudo, gerando uma lista com todas as sobreposições anotadas. Para auxiliar na decisão do “melhor” TR, ambas as ferramentas disponibilizam métricas de saída que compuseram ao menos um dos seus critérios internos de mensuração de qualidade. TRF devolve a entropia do TR em uma tentativa de estimar a complexidade da região, com menores valores para as menos complexas. É importante ressaltarmos que a entropia considerada pelas ferramentas de TRs referem-se ao conceito da

Teoria da Informação, denominado Entropia de Shannon(BENSON, 1999). Já a T-REKS gera a métrica denominada “p-sim”, que calcula o nível de similaridade entre cada parte da repetição após alinhamento múltiplo (MSA). Ela baseia-se no cálculo da distância Hamming, muito utilizada para dar pesos para substituições coluna a coluna, normalizada pelo tamanho do módulo de repetição (MR) e quantidade de repetições encontradas. Maior “p-sim” indica maior similaridade (JORDA; KAJAVA, 2009).

A decisão de quais ferramentas utilizar no processo de análise de TRs é uma tarefa difícil e trabalhosa devido a grande diversidade de ferramentas disponíveis e padrões detectados, o que faz com que geralmente apenas uma ferramenta seja utilizada (PELLEGRINI, 2015). Com o intuito de facilitar esse processo, (SCHAPER*et al.*, 2015) desenvolveram a biblioteca python TRAL, capaz de executar as ferramentas XTREAM, TRUST, T-REKS e HHrepID para AA e TRed, TRF, T-REKS para NT, quando devidamente configuradas no computador utilizado, e apresentar métricas de seleção dos melhores TRs com abordagens adicionais às apresentadas pelas ferramentas. Entre estas métricas, podemos destacar 2, geradas conforme indicação da documentação do software:

- p-valor: baseado no MSA gerado com as repetições da região do TR, que busca auxiliar na sua avaliação de significância
- “Divergência”: baseado na métrica denominada “phylo_gap01”, realiza a inserção exponencial de GAPs aplicando uma taxa de 0,1 vezes menor do que a taxa de substituições. Inserções e deleções de AAs são baseadas na distribuição Zipfian, com origens na teoria da informação.

Poucos estudos de bioinformática foram realizados nos últimos anos com o intuito de classificar e entender a composição de regiões de TR em *T. cruzi*. Eles geralmente buscam comparar homólogos entre tripanossomatídeos, como o realizado por Mendes *et al.* (2013), que teve maior enfoque em dados de TR de proteínas de superfície; e a criação de uma bsae de dados especializada em proteínas de virulência de protozoários gerada por Ramana e Gupta (2009) que efetuou uma análise muito simples de ocorrência de mono e hetero-repeats, sem grande aprofundamento.

2.2.3 Identificação de epitopos de células B

Antígenos são quaisquer estruturas moleculares capazes de causar resposta do sistema imune. Epitopos ou determinantes antigênicos são segmentos desses antígenos que, ao serem detectados, estimulam a produção dos anticorpos. As células B têm como característica o reconhecimento de regiões solventes expostas do antígeno, através de receptores de antígenos denominados receptores de células B (BCR)(PARHAM, 2009). Por essa razão os melhores epitopos devem estar contidos predominantemente em regiões de alça ou expostas da molécula (BARLOW *et al.*, 1986).

Os linfócitos T dispõem de um motivo MHC (*major histocompatibility complex*) facilmente detectável nos antígenos alvos de ferramentas, enquanto que, para linfócitos B um domínio conservado não está disponível. Para contornar esta dificuldade, as ferramentas de bioinformática desenvolvidas para predição de epitopos de célula B que tem como alvo os antígenos protéicos, contam com a predição de regiões expostas em proteínas de superfície ou na composição de aminoácidos relacionados a essa condição (SALIMI *et al.*, 2010).

O primeiro método de predição de epitopos de célula B foi publicado por Hopp e Woods (1981), utilizando propriedade dos aminoácidos através de uma escala de hidrofobicidade. Pellequer *et al.* (1991) avaliou as métricas do gênero, que utilizavam a mesma escala, a de acessibilidade ou flexibilidade, disponíveis até então, concluindo que a maioria das escalas apresentou predições pouco melhores do que a atribuição aleatória, entre 50 e 62%. Os mesmos autores posteriormente propuseram uma escala baseada em beta-*turn*, que segundo suas análises obtiveram desempenho de 80% de acertos (PELLEQUER *et al.*, 1993).

Uma nova avaliação, agora realizada por Blythe e Flower (2005) comparou os métodos baseados em propriedades únicas, concluindo uma performance espúria. A escassez de ferramentas eficazes neste período impulsionou o desenvolvimento de diversas técnicas baseadas em inteligência artificial, com o intuito de combinar características e gerar ferramentas de melhor qualidade.

As técnicas baseadas em aprendizado de máquina são atualmente as mais utilizadas, pois requerem somente a sequência como entrada e buscam predizer as regiões expostas (*coil*) da proteína. Entre elas, podemos destacar a ferramenta BepiPred (versão 2.0 -JESPERSEN *et al.*, 2017), que utiliza o algoritmo RF (*Random*

Forests Regression). Os dados utilizados para a predição são volume, hidrofobicidade, polaridade calculados para cada resíduo de aminoácido, juntamente com informações de estrutura secundária obtidas através da ferramenta NetSurfP (PETERSEN *et al.,* 2009). Adicionalmente podemos citar a ferramenta CBTope (ANSARI; RAGHAVA, 2010), que utiliza o algoritmo SVM (*Support Vector Machines*), com uma combinação de dados de características físico químicas e uma métrica denominada perfil de composição de padrões, que busca extrair características que diferenciam padrões antigênicos dos não antigênicos. Os algoritmos nos quais estas ferramentas são baseados serão explicados em mais detalhes na próxima seção.

Nenhum estudo global de regiões de epitopos de célula B foi realizado até então em *T. cruzi*. Um dos fatores que dificulta sua execução é a limitação imposta pelas ferramentas para poucas sequências por execução, devido a necessidade de cálculo da estrutura 3D da proteína, que demanda alta capacidade de processamento e tempo de execução.

2.2.4 Inteligência Artificial como ferramenta de bioinformática

Inteligência artificial (AI - *Artificial Intelligence*) tem sido cada vez mais aplicadas no desenvolvimento de algoritmos de bioinformática através de técnicas de aprendizado de máquina (ML - *Machine Learning*) (revisado por LIBBRECHT; NOBLE, 2015). As técnicas mais amplamente utilizadas com sucesso atualmente buscam automatizar o processo de tomada de decisão à partir de exemplos fornecidos no processo de aprendizagem, conhecidas como aprendizado supervisionado. Nesta competência temos a classificação, que busca atribuir uma classe para cada entrada, e a regressão, que busca predizer o valor da variável alvo.

Dispomos também de uma segunda técnica, designada aprendizado não supervisionado, com algoritmos de agrupamento. Aqui somente os dados de entrada são utilizados para tomada de decisão, sem a apresentação de exemplos. Ambos os casos partem da premissa de reconhecimento de padrões nos dados informados para tomada de decisão (MÜLLER; GUIDO, 2016).

A ML supervisionada dispõe de um processo que divide-se em 6 passos (FLACH, 2012):

1) Obtenção de dados relevantes;
2) Engenharia de atributos (*feature engineering*) ou pré-processamento;

3) Geração de um modelo;
4) Aplicação do modelo em dados de testes;
5) Avaliação da performance do modelo;
6) Melhorias e ajustes.

Após a definição do escopo do algoritmo e obtenção de dados representativos (item 1), as atividades de engenharia de atributos (item 2) buscam resolver problemas, como o tratamento e limpeza de valores faltantes ou incorretos, e efetuar transformações que tornem os atributos gerados boas representações dos dados a serem preditos ou agrupados (FLACH, 2012). Após a produção de um conjunto de variáveis relevantes, o passo seguinte é o de construção de um modelo (item 3), geralmente gerado à partir de parte dos dados disponíveis (entre 60 a 80% dependendo do volume de dados), conhecido como conjunto de treinamento. O modelo tem a função de extrair informações dos dados de entrada, gerando uma abstração do aprendizado, para que possa ser aplicada a novos dados. Ainda segundo (FLACH, 2012), o modelo pode ser baseado em transformações geométricas, que contam com conceitos matemáticos de transformação de dados; probabilísticas, relacionados com a distribuição dos dados de entrada; ou lógicas, que efetuam a divisão binária dos dados de forma a tornar as regras de segregação mais humanamente compreensíveis.

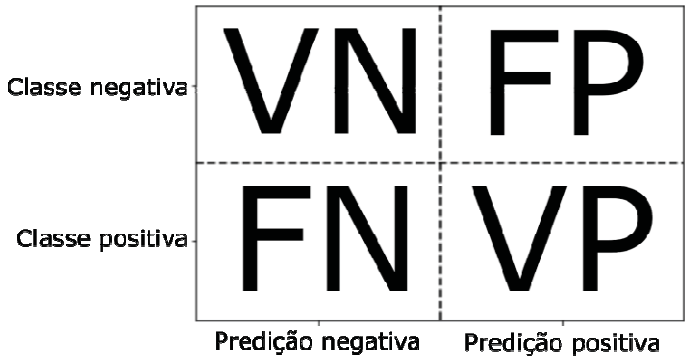
Alguns algoritmos de geração de modelos de classificação mais popularmente utilizados em pesquisas biomédicas, segundo Koohy (2017) são :

- MLP (*Multi-layer perceptron*): Principal algoritmo de redes neurais artificiais (ANN) que tenta replicar o funcionamento do cérebro humano, atribuindo os dados de entrada a neurônios interconectados. Efetua transformações lineares e confere pesos a esses dados como meio de aproximar a classe correta na camada de saída. Em etapa de geração do modelo, utiliza o algoritmo de backpropagation para propagar os erros novamente, corrigindo gradativamente as camadas de entrada e, em diversos ciclos de cálculo e correção, busca a otimização das saídas;
- SVM (*Support vector machines*): Algoritmo de classificação linear binária que tem sua definição em álgebra linear. Ele busca encontrar um plano em altas dimensões, chamado hiperplano, capaz de separar linearmente os dados, de forma a diferenciar as características das classes informadas no conjunto de treinamento;

- Random Forests: Combina o algoritmo de transformação binária árvore de decisão (Decision trees), em um conjunto de árvores randômicas, conceito conhecido como ensemble. O ensemble acumula probabilidades de acerto com o maior número de árvores e ao final a classe mais apontada entre as árvores é a escolhida.

O modelo gerado a partir de algum desses algoritmos, é então utilizado como entrada junto aos mesmos atributos usados para gerá-lo, porém referentes ao conjunto de testes (item 4). Aqui os valores ou classes a serem preditos, apesar de conhecidos, não são fornecidos ao modelo, e este tem o objetivo de predizê-los corretamente ou aproximá-los o máximo possível. Neste momento utilizamos métricas de validação de *performance* (item 5) e as principais aplicáveis para classificação binária baseiam-se na matriz de confusão (FIGURA 1) (MÜLLER; GUIDO, 2016).

FIGURA 1 - EXEMPLO DE MATRIZ DE CONFUSÃO



Fonte: Adaptado de Müller; Guido (2016)

NOTA: A matriz acima representa a contagem de ocorrências onde o preditor foi capaz de acertar corretamente as classes informadas (VN - verdadeiro negativo e VP - verdadeiro positivo, na diagonal principal) e onde o preditor classificou de maneira inversa as classes informadas (FN - falso negativo e FP - falso positivo). Esta matriz é uma das principais fontes para geração de métricas de qualidade de classificadores.

A matriz de confusão efetua a contagem das classes preditas em relação às esperadas e as seguintes métricas são geradas à partir das suas informações (MÜLLER; GUIDO, 2016):

- Acurácia: mede com que frequência o classificador está correto, tem como fórmula (VP+VN)/Total. Sua principal desvantagem é a de não considerar as classificações erradas e mascarar resultados para conjuntos desbalanceados,

isto é, onde o número de classes positivas e negativas não é igualmente distribuído;

- Precisão: esta métrica busca quantificar quantas das classificações positivas estão corretas. Sua fórmula VP/VP+FP, considera classificações erradas parcialmente;
- Sensibilidade (*recall*): Busca quantificar quantos dos casos positivos foram realmente selecionados. Sua fórmula VP/VP+FN também considera os erros parcialmente;
- F1-Score: Busca o balanço entre as métricas precisão e especificidade (2 * (precisão*especificidade)/(precisão+especificidade)), sendo menos afetada por dados desbalanceados do que a acurácia;

Finalmente, para aprimorar os resultados obtidos melhorias e ajuste (item 6) devem ser realizados. Estas podem variar de alterações nos atributos de entrada e parâmetros do modelo a meios de evitar armadilhas intrínsecas ao processo (FLACH, 2012).

Os problemas mais comuns na aplicação de técnicas de ML são o *overfitting* e predição sobre dados desbalanceados. *Overfitting* trata-se basicamente de um modelo bem ajustado, porém ineficaz para previsão de novos resultados. A principal técnica utilizada para evitar este problema é a validação cruzada (*cross-validation*) com K execuções, que seleciona amostras randômicas dos dados K vezes com o propósito de permitir uma melhor avaliação da capacidade de generalização do modelo. Devido a esta característica de múltiplas execuções ela também permite diversos testes com os parâmetros do modelo, propiciando a seleção da melhor configuração entre qualidade das predições e velocidade de execução (MÜLLER; GUIDO, 2016). O segundo problema, referente a dados desbalanceados, apresenta solução um pouco mais desafiadora. Devido a baixa disponibilidade de registros para uma das classes da amostra, o modelo apresenta dificuldade em generalizar suas características. A principal maneira de aumentar sua representatividade nas classificações consiste em balancear a base, seja através da remoção aleatória dos indivíduos da outra classe ou geração de representantes fictícios para a classe em questão. Esta técnica tem como principal custo a geração de uma grande quantidade de falsos positivos, pois o modelo torna-se mais permissivo quando apresentado a dados reais, naturalmente desbalanceados. A adoção desta estratégia deve ser sempre avaliada de acordo com o objetivo de análise. A geração

excessiva de falsos positivos pode ser prejudicial, como por exemplo em um estudo onde uma farmacêutica precisa ser informada da eficácia de uma nova droga e excesso de falsos positivos ocasionariam perdas financeiras. No entanto, em situações onde as perdas não sejam tão notórias, esta é uma opção viável por permitir ao analista dos dados tomar a decisão final de se a predição está ou não correta (MÜLLER; GUIDO, 2016).

A principal vantagem da aplicação de técnicas de inteligência artificial sobre o desenvolvimento de sistemas convencional é a de que regras específicas não precisam ser escritas para que o sistema seja capaz de operar sobre os dados e extrair os resultados esperados (MÜLLER; GUIDO, 2016). Algumas das ferramentas avaliadas por este estudo utilizam-se destas técnicas e seguem o processo descrito acima. São elas Bepipred 2.0 e CBtope, desenvolvidas para predição de epitopos de célula B e T-REKs, utilizada para anotação de TRs.

2.3 JUSTIFICATIVA

Regiões de Tandem Repeats são abundantes e de extrema importância para compreensão das características genéticas e bioquímicas do *T. cruzi*, no entanto os estudos efetuados até o momento foram direcionados para um conjunto de proteínas com características específicas e tiveram um enfoque mais comparativo com os demais tripanossomatídeos do que na análise da sua composição em si, conforme descrito na seção 2.2.2.

Desconhecemos a existência de análises exploratórias e abrangentes contemplando o papel de Tandem Repeats como epitopos de célula B pois, de acordo com nossas pesquisas, os principais antígenos mapeados para *T. cruzi* atualmente são compostos por regiões repetidas e seu mapeamento em genes nunca pesquisados pode trazer oportunidades de ampliação do conjunto de alvos sorológicos existente, de acordo com o mapeamento realizado na seção 2.1.6.

A análise de códons preferenciais já foi efetuada anteriormente em estudos envolvendo genes codificadores de proteínas de superfície, como descrito na seção 2.1.4, e as pesquisas em epitopos recombinantes atualmente exploram esta técnica como meio de geração de peptídeos de sintetização mais ágil, conforme descrito na seção 2.1.6. No entanto, não dispomos de pesquisas que busquem classificar

regiões antigênicas compostas por códons ótimos existentes no organismo, aumentando suas chances de sintetização in vivo.

Na seção 2.2 e todas suas sub-seções nós descrevemos os processos de anotação de Tandem Repeats e epitopos de célula B através de técnicas de bioinformática, que são complexos e requerem uma combinação de diversas ferramentas para que uma qualidade mínima de análise seja atingida. Ferramentas ágeis e de utilização simplificada podem favorecer as pesquisas envolvendo patógenos eucariotos e ampliar nosso escopo de conhecimento. Para doenças negligenciadas estes fatores são ainda mais significativos, pois podem representar uma redução nos custos, tempo e foco das análises.

3 MATERIAIS E MÉTODOS

3.1 PRÉ-PROCESSAMENTO DOS DADOS

Efetuamos o download dos arquivos .FASTA das sequências codificadoras de transcritos do repositório de dados TritrypDB (versão 35), navegando a interface gráfica do website pelos menus *New Search > Genes > Gene Models > Gene type*. Utilizamos os filtros *Organism* "*Trypanosoma cruzi CL Brener Esmeraldo-like"* e *gene type* "*protein coding with no pseudogenes*", totalizando 9.039 genes. Posteriormente, executamos um processo de remoção adicional dos transcritos que não atendessem aos seguintes critérios:

1. Sequências iniciadas com códon ATG;
2. Códon de terminação ao final da sequência (TAA, TAG e TGA) sem interrupções internas.

Ao final deste processo restaram 7.660 IDs/sequências, e seus "*headers*" foram usados para a seleção das sequências de proteínas também obtidas através do repositório TritrypDB.

Os dados de transcriptoma escolhidos para análise foram obtidos através do download do material suplementar do artigo publicado por Li et al. (2016) (S2_Table_tcruzi.xlsx - Aba *Raw Counts*). Sua escolha deveu-se a este ser o estudo mais abrangente publicado até a presente data, contemplando amostras de tripomastigota TCT (*tissue culture trypomastigote*), e de parasitos em infecção de celulas de fibroblasto (HFF). Estes foram induzidos à diferenciação para amastigota, com conteúdo de RNA isolado em diferentes horários pós-infecção. Adicionalmente, o RNA de epimastigotas e tripomastigotas extracelulares foi obtido por cultura axênica.

Dados de ontologia foram obtidos através do website do repositório TritrypDB com o objetivo de facilitar nossa compreensão de função de genes com e sem presença de TRs. Para cada busca em que a ontologia foi avaliada, a consulta da lista de genes foi efetuada através do menu "*New Search > Genes > Annotation, curation and identifiers > Gene (IDs)".* Na página de resultados a ferramenta apresenta o botão "*Analyze results*" que dispõe de um link denominado "*Gene Ontology Enrichment*". As três ontologias foram sempre avaliadas e o filtro padrão de corte do p-valor de 0,05 foi mantido para garantir que somente termos significantes

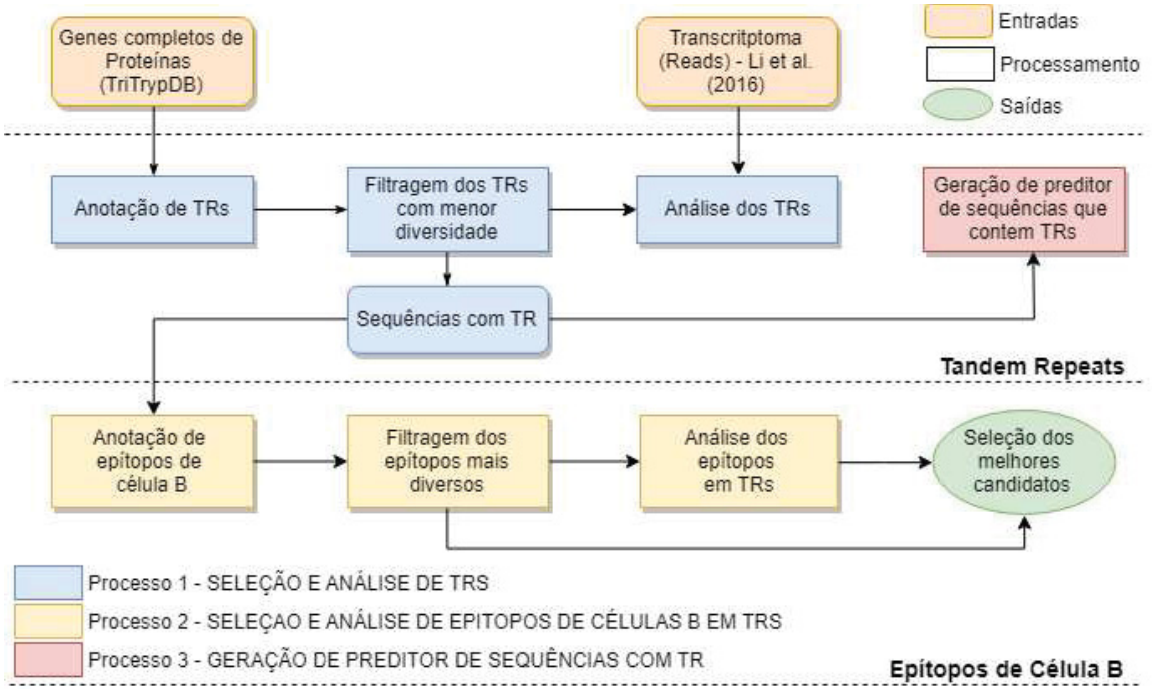
fossem mantidos e foram consideradas ontologias curadas e não curadas. O resultado de cada pesquisa foi extraído em formato .tab junto às imagens de nuvem de palavras para cada ontologia.

A relação de genes de proteínas de superfície foi também extraída do database TritrypDB, utilizando o caminho: "*New Search > Genes > Protein targeting and localization > Transmembrane Domain Count*". Os filtros padrão de mínimo e máximo de domínios foram mantidos (1 e 99, respectivamente). Somente os Gene IDs foram baixados em arquivo texto para utilização posterior.

3.2 PROCESSOS ADOTADOS

Com o intuito de realizar análises de TRs e buscar os melhores epitopos de célula B hipotéticos dentro dessas regiões, desenhamos um processo (FIGURA 2) capaz de manter o máximo de genes em escopo e, após cada uma de suas etapas, aplicamos filtros capazes de reduzi-los, mantendo os candidatos mais adequados ao final do estudo.

FIGURA 2 - FLUXOGRAMA DE ANOTAÇÃO DE TRS E EPÍTOPOS DE CÉLULAS B EM *T.CRUZI*



FONTE: A autora (2019).

Utilizando as lições aprendidas com os processo de anotação e análise de TRs, desenvolvemos um preditor de sequências que contêm TRs.

Todos os processos de tratamento das saídas das ferramentas externas e funções de apoio de análise foram implementadas utilizando a linguagem MATLAB, versão de estudante R2017B. Arquivos do tipo "*live script*" foram gerados com o intuito de registrar e facilitar a validação das saídas em cada etapa do processo. Eles permitem a execução sequencial de cada tarefa auxiliando inclusive na compreensão de todas as etapas efetuadas[4].

Os diagramas de Venn foram gerados à partir da ferramenta online disponibilizada pelo departamento de Bioinformática e Genômica Evolucionária da universidade Ghent (Bélgica)[5]. Para geração dos gráficos de análises de dados, arquivos CSV foram exportados com os dados alvo e importados através da linguagem R versão 3.4.4 e ambiente integrado de desenvolvimento RStudio versão 1.1.383, onde principalmente a biblioteca ggplot2 foi utilizada. Os gráficos da etapa de geração do preditor baseado em ML foram gerados utilizando o MATLAB.

3.2.1 Anotação e avaliação de TRs

Nosso processo de escolha das ferramentas a serem utilizadas neste estudo levou em consideração características relacionadas ao objetivo das nossas análises. Primeiramente, TRs apresentam padrões muito distintos e a execução de múltiplas ferramentas é a abordagem mais indicada (PELLEGRINI, 2015). Outro aspecto relevante para nossa pesquisa, relacionado a testes sorológicos, é o de que estudos recentes em peptídeos recombinantes baseiam-se em regiões de TRs com baixa diversidade entre suas repetições, porém com certo grau de entropia interna (SANTOS *et al.*, 2016; HERNÁNDEZ*et al.*, 2010). Ao final de nossa análise nós selecionamos algumas das ferramentas disponíveis, buscando principalmente a estratégia de detecção e extensão, mais apta a anotação de TRs com menor diversidade, conforme relação abaixo:

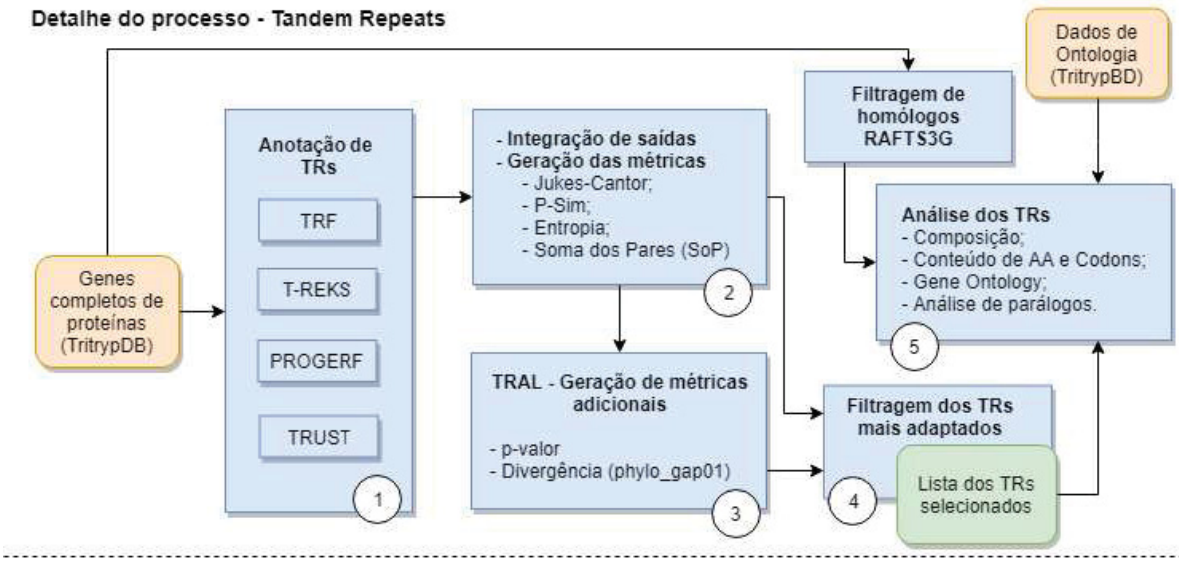
- TRF: Apesar de basear-se em sequências de nucleotídeos, esta é a ferramenta mais referenciada e de execução simples e mais rápida entre todas as selecionadas. Ajustes de fase de leitura e conversão para aminoácidos foram necessários para adequar suas anotações aos padrões das demais ferramentas;

- T-REKS: Segunda ferramenta mais utilizada, trata-se da com maior tempo de execução entre todas as selecionadas. Apresenta um volume substancialmente maior de anotações em relação às demais, aumentando a relação de possíveis candidatos;
- PROGERF: Trata-se da ferramenta desenvolvida mais recentemente para detecção de TRs baseados em dados de sequência de proteínas. Apresenta uma abordagem levemente diferente das demais e agilidade na execução;
- TRUST: Esta foi a única ferramenta selecionada com base em programação dinâmica e auto-alinhamento, justamente para avaliarmos se sua característica de maior diversidade é capaz de detectar TRs diferenciados das demais dentro de um limite de divergência interna tolerável. Apresenta velocidade de execução intermediária entre as ferramentas TRF e T-REKS.

O APÊNDICE 1 contempla a configuração detalhada utilizada em cada ferramenta anotação dos TRs.

A FIGURA 3 apresenta com mais detalhes as ações tomadas para extração e análise de TRs.

FIGURA 3 - FLUXOGRAMA DE SELEÇÃO E ANÁLISE DE TRS



FONTE: A autora (2019).

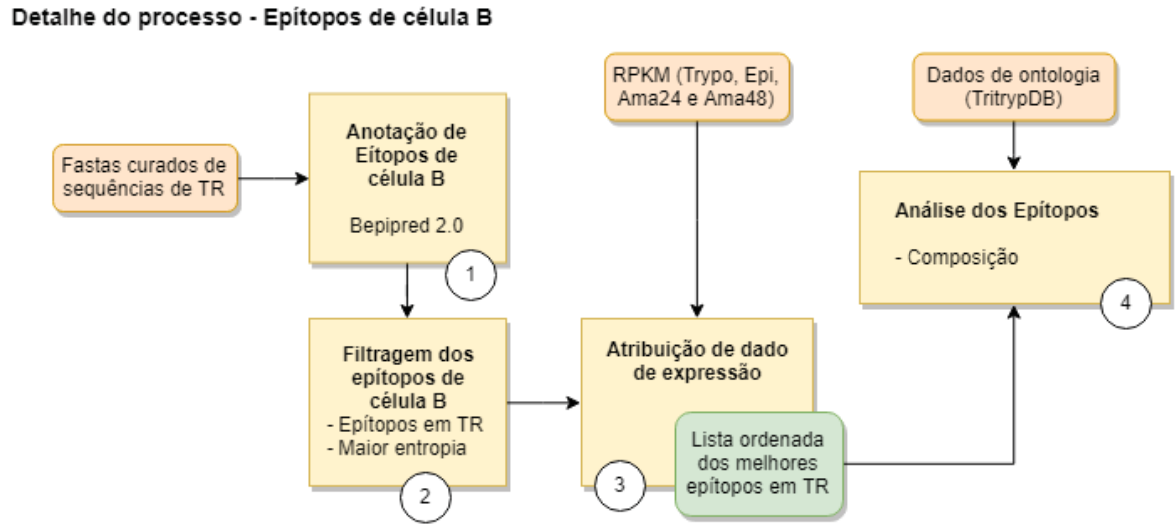
Todas as ferramentas foram executadas separadamente (1) e suas saídas, arquivos no formato texto, foram importadas e transformadas utilizando funções na linguagem MATLAB. Nesta etapa as métricas Jukes-Cantor, P-Sim, Entropia e Soma

de Pares foram geradas com base nas regiões dos TRs e dois arquivos foram gerados: um fasta com as sequências a serem anotadas e um TSV com dados mínimos dos TRs - Header da sequência, início do TR e MSA do TR (2). Ambos arquivos foram importados em uma versão adaptada da ferramenta OpenSource TRAL, desenvolvida na linguagem Python, para geração de métricas adicionais. A customização realizada permitiu a importação dos dados mínimos para execução dos cálculos das 2 métricas adicionais, p-valor e divergência, já que a funcionalidade de importação não estava disponível até a data da execução desta tarefa (3). Esta lista foi então exportada e o processo de filtragem dos TRs foi executado em arquivo *"live script"* do MATLAB (4). A lista curada de TRs agora estava disponível para anotação de epitopos de célula B e análises exploratórias adicionais (5).

3.2.2 Anotação e avaliação de epitopos de célula B

A FIGURA 4 apresenta o processo detalhado de anotação de epitopos de célula B em TRs.

FIGURA 4 - FLUXOGRAMA DE SELEÇÃO E ANÁLISE DE EPÍTOPOS DE CÉLULA B EM TRS



FONTE: A autora (2019).

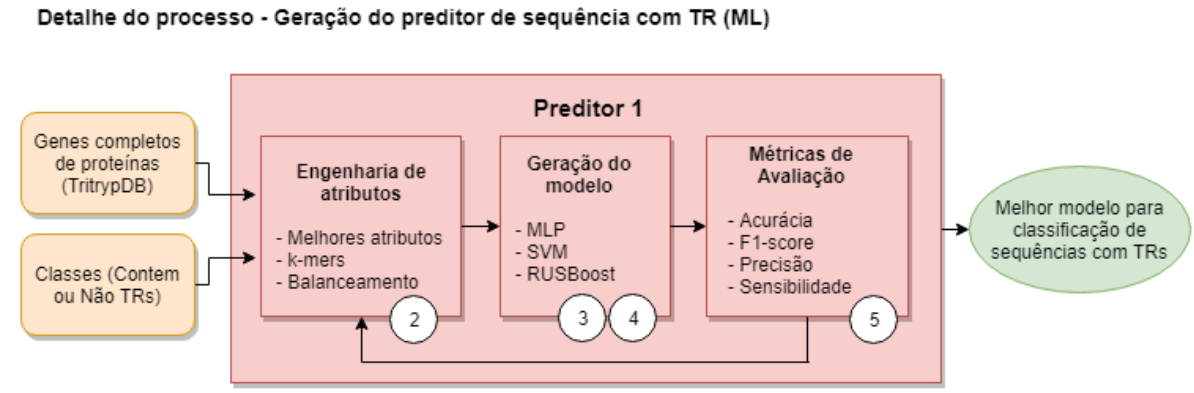
A lista de TRs selecionada foi importada através de múltiplos arquivos fasta com até 50 genes cada na ferramenta online Bepipred 2.0, que foi selecionada por tratar-se da mais referenciada para este tipo de análise (1). Ela não exige qualquer

parametrização adicional e, ao final de cada execução, a exportação de um arquivo CSV fica disponível. As saídas geradas foram importadas por função criada na linguagem MATLAB, e nesta etapa filtragem dos epitopos em regiões de TR com maior entropia foi executada (2). Valores de RPKM originários do transcriptoma para diferentes estágios do ciclo de vida foram calculados e utilizados para ordenação da lista epitopos em TR melhor adaptados (3). Informações semelhantes às da lista global de TRs foram extraídas somente dos epitopos selecionados para fins comparativos (4).

3.2.3 Geração de máquina de aprendizado de padrões de TRs

A FIGURA 5 demonstra o processo detalhado adotado para criação do modelo baseado em ML, que tem o objetivo de identificar sequências candidatas a conter TRs para submissão posterior a ferramentas de anotação de epitopos de célula B. O processo de ML adotado contem 5 dos 6 passos descritos na seção 2.2.4 deste documento.

FIGURA 5 - PROCESSO DETALHADO DE GERAÇÃO DE PREDITOR DE TRS



FONTE: A autora (2019).

Temos como entrada referente ao item (1), obtenção de dados relevantes, somente o fasta de sequências válidas de proteínas do organismo alvo da análise. Para o propósito de treinamento do modelo precisamos de um vetor adicional com a informação de se aquela sequência contém TRs ou não, o que denominamos classe

0 (sem TR) ou 1 (com TR). Aqui não nos preocupamos com quantidades, somente com a existência de TRs.

Para a etapa seguinte (2), efetuamos diversas transformações em cada sequência de entrada com o objetivo de extrairmos informações matemáticas dos genes. Todos os algoritmos testados exigem dados numéricos, portanto nosso objetivo foi o de representar com valores as diferenças e semelhanças entre cada sequência. O processo de definição dos atributos selecionados está descrito na seção 4.2.1 deste documento, relacionada aos resultados para aplicação de técnicas de IA.

O passo seguinte do processo é o (3) de geração de um modelo. Os 3 algoritmos citados na seção 2.2.4 foram utilizados com parâmetros iniciais sugeridos pela documentação da linguagem MATLAB.

- **MLP:** Algoritmo requer a informação do número de "neurônios" a serem utilizados e quantidade de camadas internas. Iniciamos os testes com 1 camada contendo 5 neurônios, pois uma quantidade muito grande pode deixar o algoritmo muito lento e não necessariamente melhorar a qualidade das predições. Entre as funções de treinamento disponíveis atualmente para realização dos cálculos nas camadas, a padrão, denominada Levenberg-Marquardt, foi utilizada por apresentar melhor performance e taxa de acerto;
- **SVM:** A função interna ou "de *kernel*", selecionada para nossos testes chama-se RBF e costuma apresentar melhor performance do que a padrão, linear;
- **Random Forests (RF):** São diversos os métodos de agregação disponíveis para algoritmos "*ensemble*". Segundo a documentação da linguagem, a função "RUSBoost" apresenta melhor performance com relação a dados com classes desbalanceadas do que a função padrão "Bag".

A próxima etapa é a de utilização do modelo (4). Adotamos a estratégia de aplicação no conjunto de treinamento e no de testes, pois esta prática permite a verificação rápida de "*overfitting*". As métricas de avaliação de performance (5) utilizadas para avaliar a matriz de confusão gerada por cada teste nesta etapa foram acurácia e especificidade.

É importante ressaltar que este processo apresenta iterações paralelas entre cada etapa, sendo necessária a execução de todos os passos em sequência diversas vezes para geração de modelo capaz de efetuar classificações de forma satisfatória.

4 RESULTADOS

4.1 ANÁLISE DE DADOS

4.1.1 Transformação inicial de dados

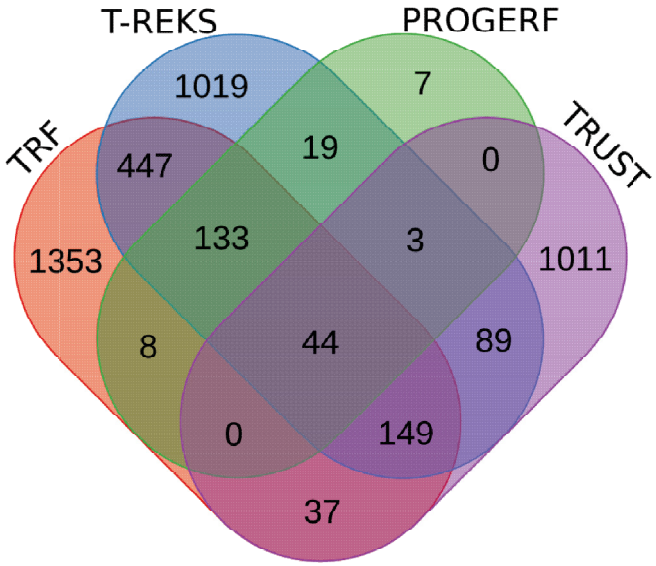
A partir de um proteoma composto por 10.338 proteínas codificadas/preditas para o haplótipo *T. cruzi* CLBrener Esmeraldo-like, foi feito um processo de curagem gerando 7.660 proteínas caracterizados como “proteínas completas”. Arquivos fasta distintos com sequências codificadoras e de aminoácidos, contemplando somente os genes selecionados foram gerados. Todo este conjunto de dados foi submetido às ferramentas de detecção de TRs, o que permitiria análises adicionais com informações de regiões repetidas.

4.1.2 Diferentes ferramentas divergem na detecção de padrões repetitivo diversos

Ao analisarmos os resultados obtidos à partir das ferramentas de anotação de *tandem repeats* percebemos que as ferramentas T-REKS e TRF apresentaram o maior número de anotações devido ao grande número de sobreposições geradas internamente, 4.622 e 1.444 respectivamente, seguidas pelas ferramenta TRUST e PROGERF, sem sobreposições. Um total de 12.021 TRs foram detectados para 3.549 sequências. Já nesta etapa inicial da análise decidimos utilizar o critério de p-valor gerado pela ferramenta TRAL para eliminar TRs com alinhamentos não significativos entre suas partes. Verificamos essa necessidade na ação de correção de fase de leitura e tradução dos TRs em AA para as saídas da ferramenta TRF, pois nem todo TR significativo em NT mostrou-se igualmente significativo em AA. Regiões com p-valor > 0,05 foram removidos das análises seguintes, restando um total de 10.435 TRs para 3.040 sequências.

No diagrama de Venn da FIGURA 6 podemos analisar a distribuição dos TRs significativos entre as ferramentas. As ferramentas TRF, TRUST e T-REKS apresentaram um grande número de TRs únicos, enquanto a PROGERF não foi capaz de predizer uma boa quantidade de TRs em geral, com apenas 7 TRs únicos (3%) e os demais sobrepostos às anotações das outras ferramentas. A TABELA 1 sintetiza as anotações por ferramenta, bem como sua contagem por gene.

FIGURA 6 - DIAGRAMA DE VENN COM A DISTRIBUIÇÃO DOS TRS ENTRE AS 4 FERRAMENTAS AVALIADAS



FONTE: A autora (2018).

TABELA 1 - CONTAGEM DE ANOTAÇÃO DE TRS POR FERRAMENTA

Ferramentas	Genes	TRs únicos	Total de TRs
PROGERF	209	7	235
T-REKS	1.868	1.019	5.718
TRF	2.004	1.353	3.119
TRUST	1.309	1.011	1.363

FONTE: A autora (2018).

NOTA: O total dos genes por ferramenta é superior ao real devido às anotações justapostas e possibilidade de ocorrência de mais de um TR em uma mesma sequência. O total de TRs apresentado já desconsidera os TRs com p-valor muito alto pois seu nível de diversidade interna os descaracteriza como TR.

Um aspecto comum a todas as análises de TR que dificulta enormemente a tarefa de verificação de “qualidade” é a definição do que é uma repetição real. Divergências simples como a da posição do aminoácido onde ele se inicia ou termina na sequência já dificultam sua avaliação, conforme exemplo do QUADRO 1. O mesmo quadro demonstra um cenário comum de sobreposição de TRs com diferentes tamanhos detectados pela mesma ou por várias ferramentas, sendo necessária a redução de redundância.

QUADRO 1 - DETECÇÃO DE TRS POR MÚLTIPLAS FERRAMENTAS

Ferramenta	Posição	TR	MSA
TRF	10-73	AKPAAKTAAKTAAKPAAKSAAKPAA KPAAKPAAKPAAKPAAKTAAKPAKK PAVKPTVKPAAKx	AKPA AKTA AKTA AKPA AKSA AKPA AKPA AKPA AKPA AKPA AKTA AKPA KKPA VKPT VKPA AK- -
T-REKs	17-60	AAKPAAKTAAKTAAKPAAKSAAKPA AKPAAKPAAKPAAKPAAKTAAKPAK KPAVKPTVKP	AAKP AAKT AAKT AAKP AAKS AAKP AAKP AAKP AAKP AAKP AAKT AAKP AKKP AVKP

FONTE: A autora (2018).

NOTA: Mesmo TR detectado pelas ferramentas TRF e T-REKs para a sequência TcCLB.506401.320 (Proteína ribossômica L7a) apresenta iníco e final diferentes.

Buscamos então na literatura alguns métodos utilizados para esse tipo de seleção. Mendes *et al.* (2013) e Richard *et al.* (2016) analisaram TRs com objetivos diversos e ao tentar utilizar esses estudos como referência nos deparamos com métricas que logo percebemos não atender nossas necessidades. O primeiro utilizou uma ferramenta que não gera sobreposições, logo indicaria automaticamente os TRs ideais, o que já foi descartado na decisão de análise com várias ferramentas. O último, voltado a identificar TRs com função estrutural, utilizou os critérios de quantidade de GAPs, tamanho da repetição e do módulo de repetição (MR), e ao tentar aplicar os mesmos parâmetros nos deparamos com uma seleção muito diversa, com muitas substituições e inserções de GAPs.

Decidimos então efetuar uma sucessão de análises de métricas sugeridas pelas ferramentas de TR, que nos permitissem definir critérios simples de seleção baseados em análises empíricas. Verificamos individualmente as seguintes métricas:

- “Jukes-Cantor” cálculo de distância sugerido pelo MATLAB (menor - melhor);
- “Divergência” (DIV) da ferramenta TRAL (menor - melhor);
- “P-Sim” da ferramenta T-REKs (maior - melhor);
- “Soma dos Pares” (SoP), métrica para avaliação de MSAs com substituição - (maior-melhor).

Destas, a que apresentou-se como melhor critério de seleção inicial foi DIV, que a um valor inferior a 0,4, mostrou-se capaz de penalizar adequadamente a inserção de GAPs e substituição de AAs. Portanto, para a seleção da melhor sobreposição, quando existente, a menor DIV seria escolhida. No entanto verificamos que para TRs mais longos que contivessem TRs internos menores, esta métrica selecionava os mais curtos e menos diversos, o que faria com que um volume grande da região dos TRs fosse perdida. Por isso, seguimos com uma iniciativa de combinação das métricas capaz de balancear a penalização entre tamanho dos TR e diversidade. Notamos que a métrica “Jukes-Cantor” é mais restritiva do que a DIV por efetuar o alinhamento do MSA com o módulo de repetição do TR ao invés de somente entre suas partes. Verificamos que esta métrica poderia ser utilizada com valores inferiores a 0,32 e que a de maior SoP deveria ser adotada como critério de seleção. Para TRs curtos SoP e DIV coincidem. O QUADRO 2 exemplifica os cenário de seleção.

QUADRO 2: SOBREPOSIÇÕES DE TRS PARA UM GENE

#	Posição	Consenso	TR	MSA	Div	Jukes-Cantor	Soma dos Pares
a	39 - 111	GWWGG GGGGG GG	GWGSDGNAGGGGG WGSGGSGGGGGGW GSGGGGGGSRGGW GSGGSGGSGGGWGS GGSGSGGGGWGSG GGGGGRG	GWGSDGNAGGG-G GWGSGGSGGGG-G GWGSGGGGGGSRG GWGSGGSGGSG-G GWGSGGSGSGG-G GWGSGGGGGGR-G	0,226	0,872	112,5
b	47 - 56	G	GGGGGGGGGG	G G G G G G G G G G	0,000	0,000	22,5
b	47 - 56	G	GGGGGGGGGG		0,000	0,000	22,5
	47 - 56	GGG	GGGGGGGGGG	GGG GGG GGG G - -	0,197	0,106	12,5
c	69 - 125	RGGWG GGGSG GGGGG WGSGG GGGGG	RGGWGSGGSGGGG GGWGSGGGGGGSR GGWGSGGSGGSGG GWGSGGSGSGGGG WGSGG	'RGGWGSGGSGGGG GGWGSGGGGGGS RGGWGSGGSGGSG GGWGSGGSGSGG -GGWGSGG---------- ----	0,168	0,261	55,0
	100 - 107	GGS	GGSGGSGG	GGS GGS GG -	0,176	0,118	7,5

FONTE: A autora (2019).

NOTA: TRs anotados para o gene TCCLB.503419.50. Para fins explicativos somente 6 sobreposições foram mantidas. O item a) linha em vermelho, demonstra qual seria a seleção somente pelo critério “DIV inferior a 0,4”, porém o critério “Jukes-Cantor” iria eliminá-la. Já o item b), com linhas em amarelo, representa a menor DIV, onde TRs muito curtos e pouco diversos seriam selecionados. Por fim, o item c), linha em verde, representa a seleção utilizando o conjunto completo de critérios, a maior SoP seria selecionada, apresentando um TR maior e suficientemente diverso.

Em resumo, os critérios de seleção dos TRs mais adequados foram:

1. Os TRs com Div <= 0,4 e Jukes-Cantor <= 0,32 seguiriam para a próxima etapa de avaliação;
2. Os de maior SoP seriam selecionados;

3. Em caso de ainda restarem sobreposições devido a empate de valores de SoP, os de menor DIV seriam mantidos;
4. Ao final, restando somente os TRs idênticos, o primeiro apresentado seria o selecionado.

A planilha total de TRs com uma coluna adicional contendo qual a etapa da sua remoção ou indicação de se foi mantido pode ser encontrada no Material Suplementar, aba “lista_total_trs”. Os critérios “P-sim” e “Entropia” foram mantidos para informação, porém não foram utilizados como parâmetro de análise na seleção dos TRs por não apresentarem nenhuma informação adicional aos critérios escolhidos.

O processo de limpeza gerou um total de 2.161 TRs para 1.680 sequências. O estudo efetuado por Mendes *et al.* (2013), utilizou o proteoma completo contendo os 2 haplótipos de *T. cruzi*, com 16.276 sequências válidas, selecionadas com os mesmos critérios de filtragem de genes válidos deste estudo. Este avaliou que quase 38% do proteoma continha regiões repetidas, enquanto nosso estudo apresenta um resultado mais conservador, de 22%. Como o estudo anterior foi realizado com a ferramenta predecessora da PROGERF e os detalhes dos TRs não foram disponibilizados, não temos meios de precisar a real causa da divergência.

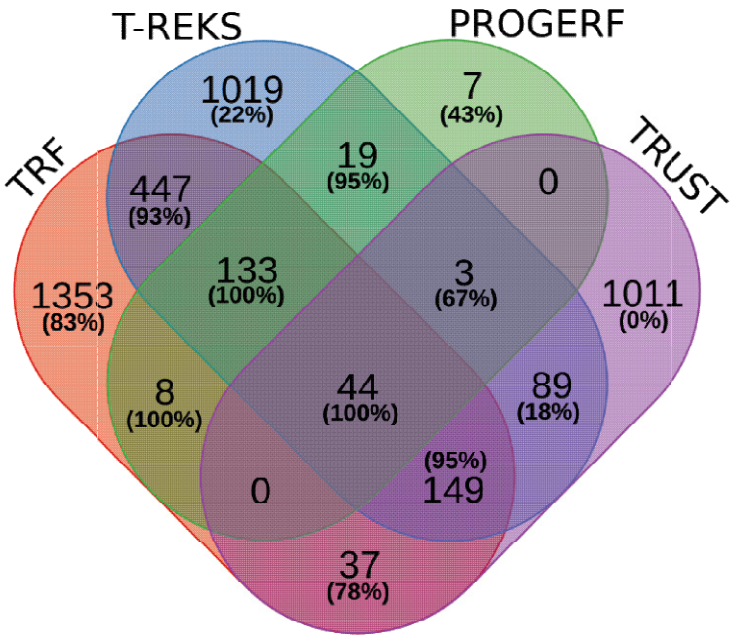
A TABELA 2 apresenta a contagem geral dos TRs por ferramenta, enquanto que o diagrama de Venn (FIGURA 7) demonstra os percentuais de TRs selecionados para cada intersecção dos conjuntos.

TABELA 2 - CONTAGEM DOS TRS SELECIONADOS POR FERRAMENTA

Ferramenta	Qtde de TRs	Qtde de Genes
TRF	1477	1209
T-REKS	655	608
PROGERF	29	29
TRUST	0	0
TOTAL	**2161**	**1846**

FONTE: A autora (2019).

FIGURA 7 - DIAGRAMA DE VENN COM PERCENTUAIS DE SELEÇÃO POR INTERSECÇÃO



FONTE: A autora (2019).

NOTA: Diagrama de Venn contém percentuais de TRs selecionados para cada intersecção entre ferramentas. É importante ressaltar que somente 1 TR foi escolhido em cada caso. Esta representação não permite a identificação de qual ferramenta teve o TR selecionado, porém demonstra a visão de que geralmente as intersecções entre as demais ferramentas com a TRF tendem a ter um maior número de TRs mantidos.

Podemos observar que as intersecções mais centrais, que envolvem sobreposições de 4 ou 3 das ferramentas, tiveram boa parte dos TRs selecionados, o que indica um menor nível de divergência entre os integrantes desses conjuntos. Todas as intersecções envolvendo a ferramenta TRF apresentaram um alto percentual de seleção, como já era esperado devido sua característica de TRs menos diversos. Já para os casos de anotação única das demais ferramentas, o percentual de seleção foi baixo, chegando a total eliminação dos TRs anotados exclusivamente pela ferramenta TRUST. Aqui temos indício de TRs mais divergentes.

Para entendermos quais as ferramentas com seleção efetiva dentro das intersecções, geramos a QUADRO 3, que apresenta uma distribuição dos percentuais para cada uma delas.

QUADRO 3 - PERCENTUAIS DE DETECÇÃO POR FERRAMENTA PARA CADA INTERSECÇÃO DO DIAGRAMA DE VENN

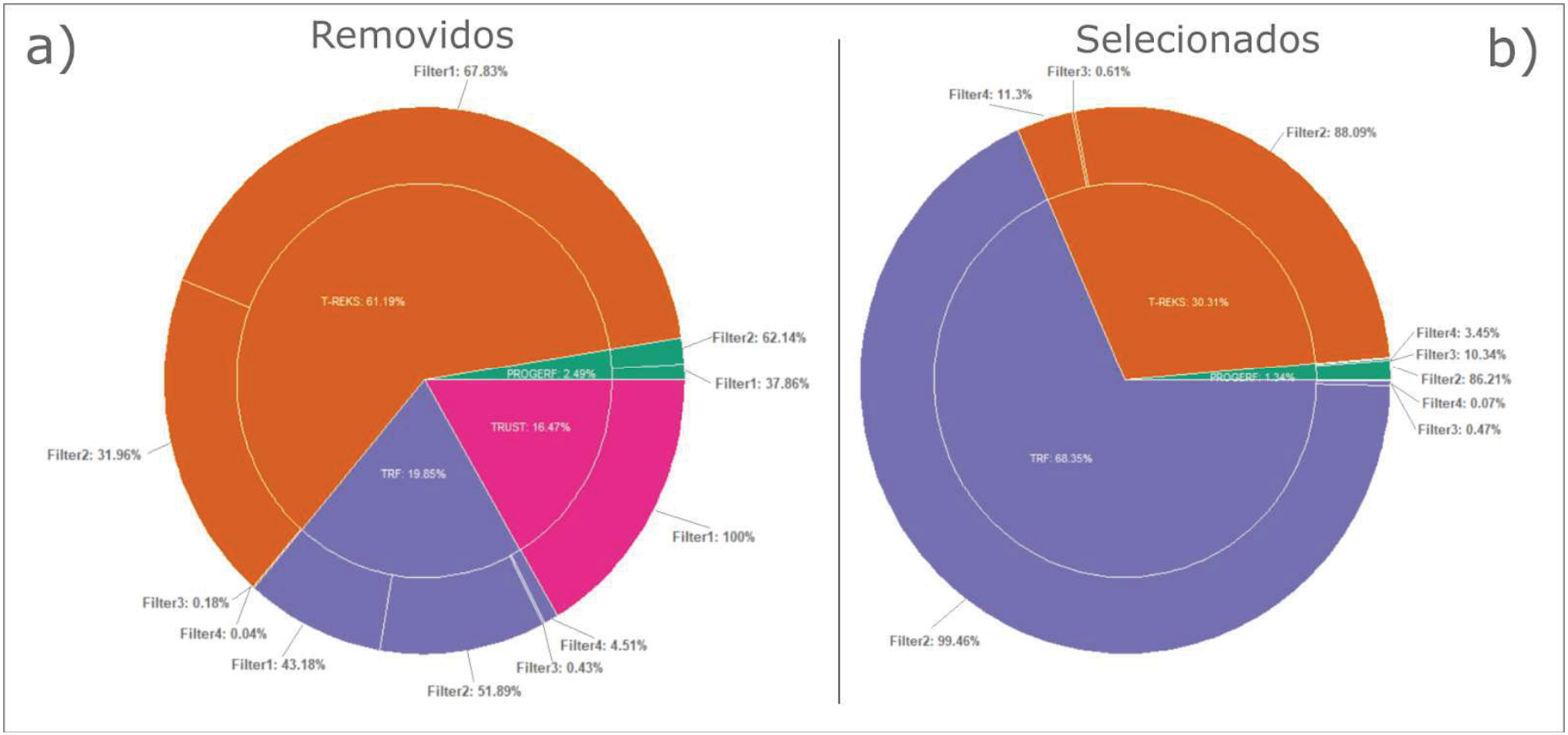
#	Intersecção	Qtde Inicial	Qtde Final	TRF	T-REKS	PROGERF	TRUST
1	PROGERF+T-REKS+TRF+TRUST	44	44	52,27%	43,18%	4,55%	0,00%
2	PROGERF+T-REKS+TRF	133	133	26,32%	64,66%	9,02%	
3	T-REKS+TRF+TRUST	149	141	36,17%	63,83%		0,00%
4	PROGERF+T-REKS	3	2		63,83%	36,17%	
5	T-REKS+TRF	447	414	40,58%	59,42%		
6	PROGERF+TRF	8	8	75,00%		25,00%	
7	TRF+TRUST	37	29	100,00%			0,00%
8	PROGERF+T-REKS	19	18		50,00%	50,00%	
9	T-REKS+TRUST	89	16		100,00%		0,00%

FONTE: A autora (2019).

Podemos verificar que um grande número de TRs anotados pela ferramenta T-REKS foi selecionada, por vezes superando as seleções da ferramenta TRF (Intersecções numeradas no QUADRO 3 como 2, 3, 4 e 5), o que pode indicar um melhor posicionamento com relação ao início e final do TR em relação às demais ou mesmo melhor distribuição de tamanho e diversidade.

Em uma análise final do processo de seleção, visualizamos a distribuição dos TRs eliminados por etapa de filtragem, considerando o Filtro 1 como o principal responsável pela eliminação de TRs muito diversos e os 3 filtros seguintes como critérios de desempate entre sobreposições. O GRÁFICO 1 apresenta essa distribuição para cada ferramenta.

GRÁFICO 1 - DISTRIBUIÇÃO DAS ETAPAS DE FILTRAGEM POR FERRAMENTA



FONTE: A autora (2019).

NOTA: Gráficos de setores em 2 níveis representando as ferramentas e em qual etapas de filtragem a) os TRs foram removidos e b) os TRs restantes foram selecionados.

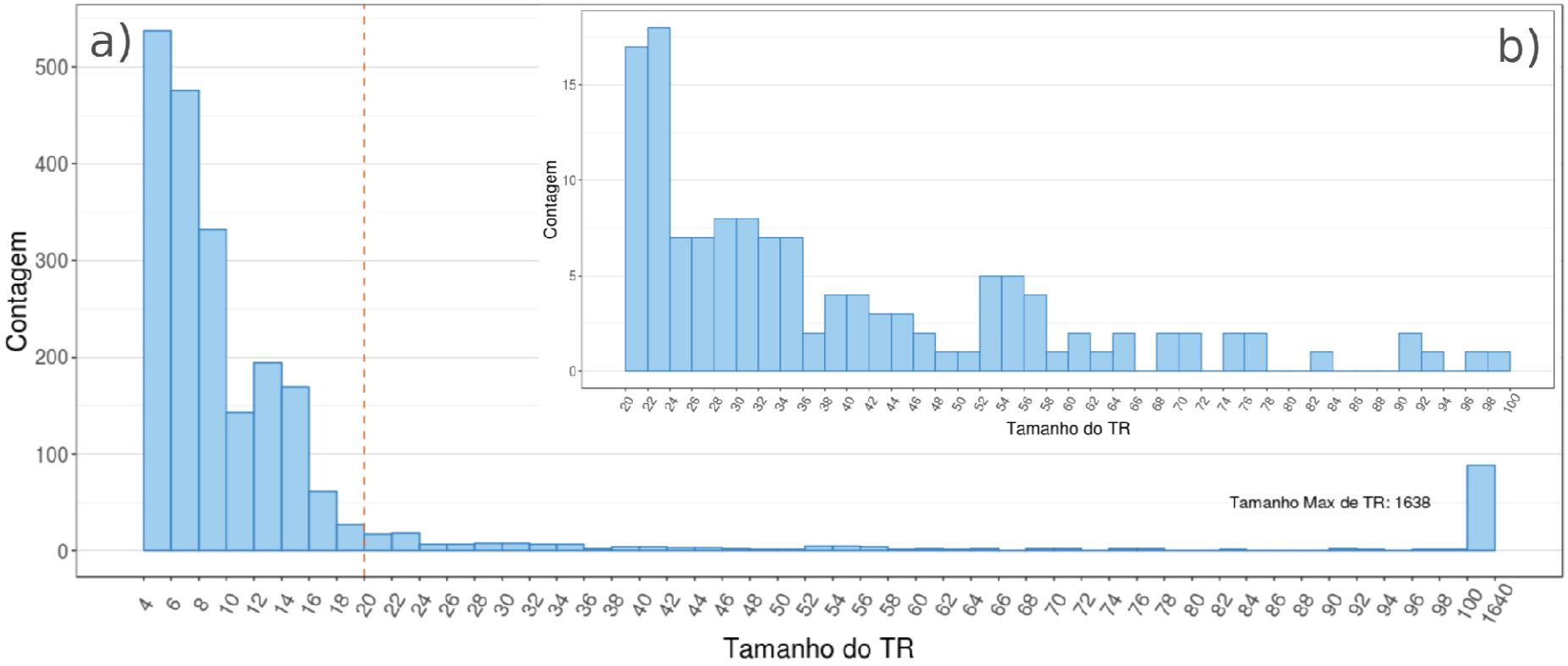
Conforme indícios já visualizados na FIGURA 7, e GRÁFICO 1 (a) confirma que todos os TRs anotados pela ferramenta TRUST eram muito diversos para passar para o nível 2 de filtragem, bem como quase 68% dos TRs anotados pela ferramenta T-REKS. Esta sofreu muito mais eliminações do que as demais devido ao grande número de anotações sobrepostas geradas. Na GRÁFICO 1 (b) verificamos que a etapa 2 de filtragem foi responsável pela manutenção de 88% dos TRs da T-REKS, indicando bons candidatos, mesmo em relação aos anotados pela ferramenta TRF, que nesta etapa teve o maior número de seleções. As 2 etapas seguintes foram responsáveis basicamente pelo desempate entre TRs muito semelhantes.

Podemos constatar que, para os parâmetros adotados em nossa pesquisa, ambas as ferramentas TRF e T-REKS são indispensáveis para a seleção de TRs adequados, corroborando com a afirmação de Pellegrini (2015) de que a utilização do arsenal de algoritmos disponíveis é importante para maior sucesso das análises. A riqueza de funções e tipos de padrões deve ser sempre considerada, permitindo ao pesquisador avaliar a quantidade de esforço e tempo atribuídos a essa tarefa de seleção das ferramentas mais adequadas e tratamento dos diversos dados de saída apresentados. Os critérios selecionados para o processo de filtragem foram capazes de representar matematicamente as expectativas de análise e podem ser bons parâmetros para este tipo de análise.

4.1.3 Genes de *T. cruzi* contém TRs predominantemente curtos

Os TRs mapeados para um baixo nível de diversidade são majoritariamente curtos. O GRÁFICO 2 apresenta sua distribuição por tamanho, com a demarcação pela linha tracejada vermelha representando 93% dos dados para TRs até 20 AAs.

GRÁFICO 2 - HISTOGRAMAS COM CONTAGEM DE TRS POR TAMANHO



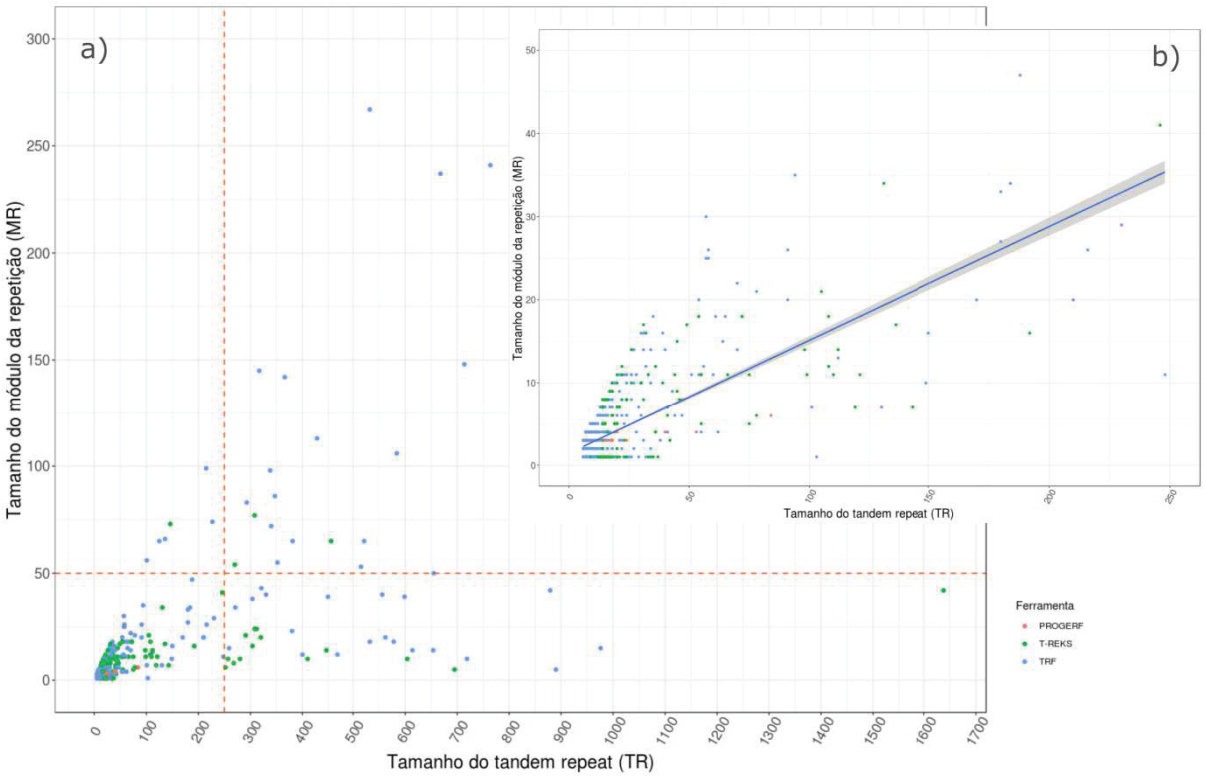
FONTE: A autora (2019)

NOTA: O histograma a) representa a contagem de TRs por tamanho. A linhas tracejada vermelha delimita tamanhos até 20 AAs, totalizando 93% dos TRs. A área à direita da linha é representada em detalhe no histograma b) entre 20 e 100 AAs, totalizando 134 TRs (7%). A última coluna do gráfico agrupa todos os TRs com tamanho entre 100 e 1.640 AAs, totalizando 87 TRs (4%).

Avaliando a relação entre tamanho total dos TRs e de seus módulos de repetição verificamos uma relação média de quase 7x (GRÁFICO 3 - b). O gráfico de bolhas (GRÁFICO 3 - a) nos auxilia na percepção da área de todos os TRs. As cores diferenciam as ferramentas que os detectaram, sem relação perceptível com os tamanhos apresentados.

Estudos voltados para análise estrutural de regiões do gene com TRs indicam que quando estes são maiores do que 50 AAs, são capazes de dobrar-se de forma independente, gerando regiões de domínio estáveis (revisado por KAJAVA, 2012). Nosso processo de seleção detectou 123 TRs com essa característica (17,2%).

GRÁFICO 3 - ÁREA DOS TR E SEUS MÓDULOS DE REPETIÇÃO



FONTE: A autora (2019).

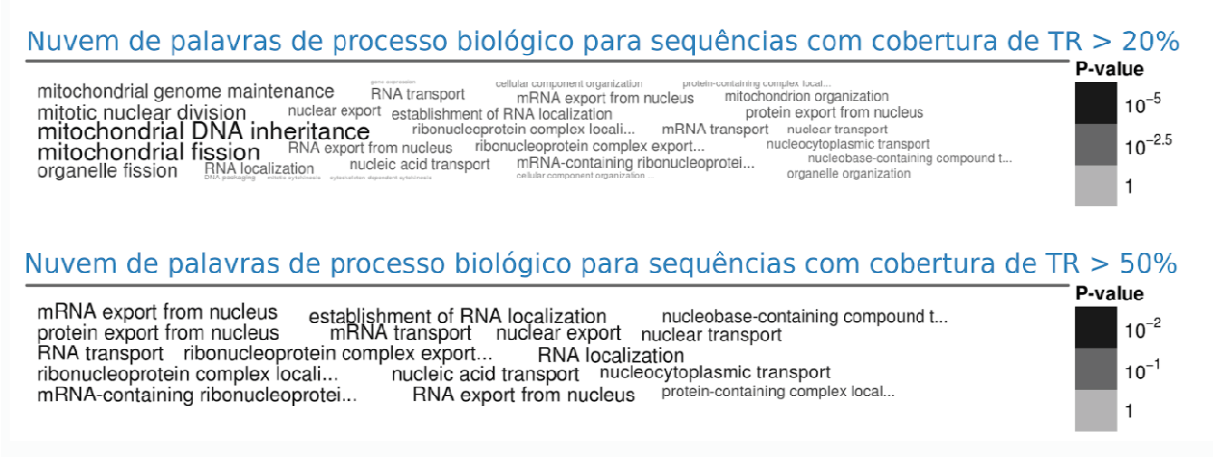
NOTA: O gráfico de bolhas a) representa a relação entre tamanho do TR e tamanho do MR. A linhas tracejadas vermelhas delimitam 99,97% dos TRs (2.101), que limitam-se a área total de 250 AAs e módulos de repetição de no máximo 50 AAs. O gráfico b) apresentam uma visualização mais detalhada desta região, com linha de tendência apresentando uma relação entre tamanho do TR e do MR de quase 7x.

Observamos a representatividade de TRs degenerados, que são os que apresentam ao menos 2 aminoácidos distintos em toda sua extensão e obervamos 751 (34,75%). Essa divergência deve-se provavelmente a grande quantidade de TRs diversos removidos em nosso estudo, tornando os TRs com menor entropia mais significativos na conjunto.

Buscamos também caracterizar a relevância dos TRs relativos à sequência completa. Podemos visualizar no GRÁFICO 4 que 93,9% das sequências têm até 20% da sua área total coberta por TRs (1.577 genes). Aqui as sequências com múltiplos TRs tiveram suas áreas agregadas para que a região real de cobertura fosse avaliada.

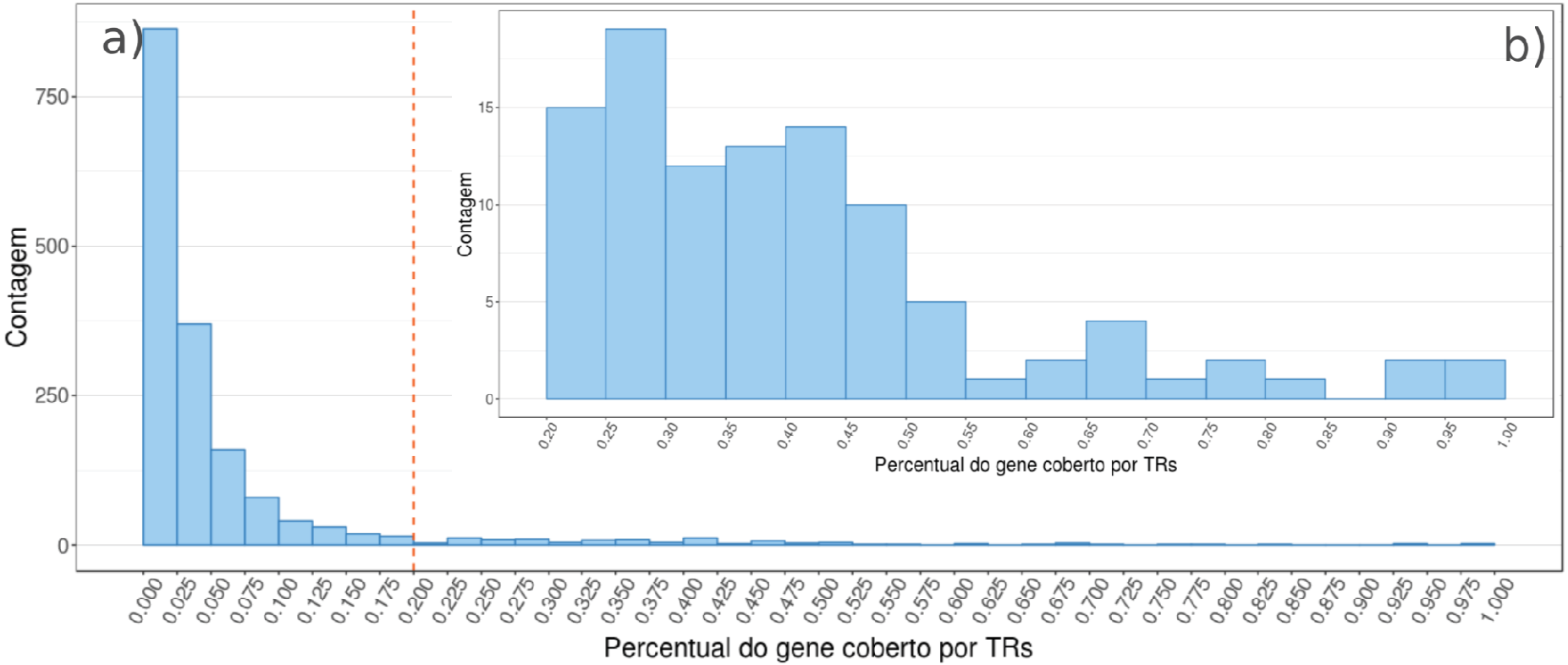
A lista de 103 sequências com área relativa superior a 20% é composta por 38 genes de proteínas hipotéticas (36,9%). Ao submetermos sua relação à análise de enriquecimento de GO obtivemos um grupo de termos de processo biológico levemente mais associados a funções mitocondriais, enquanto que para cobertura superior a 50% sobressaíram-se os termos relacionados aos processos de metabolismo de RNA (FIGURA 8).

FIGURA 8 - NUVEM DE PALAVRAS DE GO PARA TRS COM COBERTURA RELEVANTE EM SEUS GENES DE ORIGEM



FONTE: A autora (2019).

GRÁFICO 4 - HISTOGRAMA DA RELAÇÃO ENTRE TAMANHO DA SEQUÊNCIA E DA REGIÃO DE COBERTURA DE TRS



FONTE: A autora (2019).

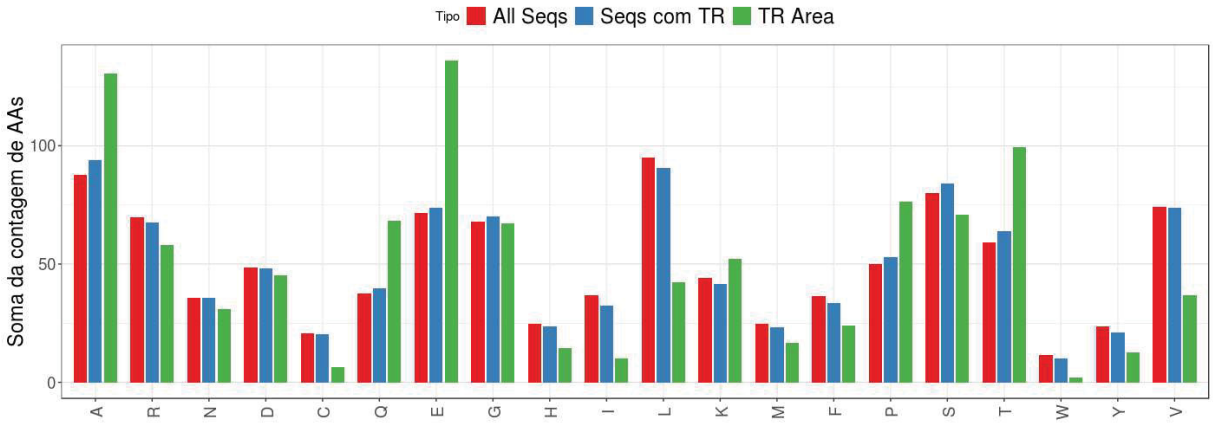
NOTA: O histograma a) representa a contagem de TRs pelo percentual que estes ocupam no gene de origem. A linhas tracejada vermelha delimita até 20% do tamanho, totalizando 93,9% das sequências com TR (1.577). A área à direita da linha é representada em detalhe no histograma b.

Estas características podem indicar uma tendência a TRs mais longos terem funções estruturais para estas classes de proteínas. No entanto, é importante observarmos que as anotações de GO são mais ricas para essas famílias pois elas são muito estudadas, o que pode ocasionar um viés em análises de ontologia do *T. cruzi*.

4.1.4 Preferências de AA são distintas entre TRs e sequências completas

Seguimos então com a verificação da composição de códons e AAs com o objetivo de avaliar se as regiões de TR seguem o padrão detectado em todo o conjunto de genes. Para esta etapa a contagem AAs foi efetuada em todo o conjunto das 7.660 sequências, nas 1.680 sequências com TRs e nas áreas dos 2.161 TRs (GRÁFICO 5). Podemos notar no GRÁFICO 5 que a distribuição dos AAs para as sequências de TR com relação a todo conjunto de sequências válidas seguem a mesma tendência, porém as regiões de TR apresentam algumas preferências distintas por alanina (A), ácido glutâmico (E), treonina (T), prolina (P) e glutamina (Q), nesta ordem. Leucina (L) e valina (V), bastante representativas em sequências completas, apresentaram menor ocorrência em regiões de TR, enquanto que os já incomuns triptofano (W), isoleucina (I), cisteína (C) e histidina (H) perdem ainda mais representatividade nessas regiões.

GRÁFICO 5 - COMPARAÇÃO DE COMPOSIÇÃO DE AAS ENTRE TODAS AS SEQUÊNCIAS E TRS



FONTE: A autora (2019).

NOTA: Gráfico normalizado em frequências por 1.000 traz todos os conjuntos de dados a uma escala comparável.

Para analisarmos as regiões de TR, decidimos primeiramente avaliar o efeito gerado por proteínas com alta homologia sobre o conjunto de TRs como um todo. Para isso, rodamos a ferramenta RAFTS3G buscando parálogos com 70% de identidade e avaliamos a ocorrência de genes de superfície entre eles. A busca de homólogos em todo o conjunto de genes poderia vincular ao mesmo cluster genes com TR com baixa identidade entre si, mas com alta identidade nas regiões externas ao TR. Buscando amenizar este efeito, optamos por efetuar a análise de homologia somente entre os 1.680 genes com TRs. Detectamos 450 genes parálogos (26,78%) e destes 206 são transmembrana (45,78%) segundo relação obtida através do banco de dados TriTrypDB.

Como proteínas de superfície são abundantes no genoma de *T. cruzi* e a presença de TRs em sua composição é conhecida (BUSCAGLIA et al., 2006), decidimos realizar análises também das sequências de TR com estas características. Entendemos que quaisquer variações detectadas na lista sem os homólogos seria compensada na lista que contempla somente os genes com domínio transmembrana devido a sua predominância na relação de parálogos, porém é importante ressaltar que o conjunto sem parálogos contempla genes de superfície.

A TABELA 3 apresenta um resumo dos percentuais obtidos considerando TRs em proteínas de superfície e demais proteínas para TRs degenerados e perfeitos.

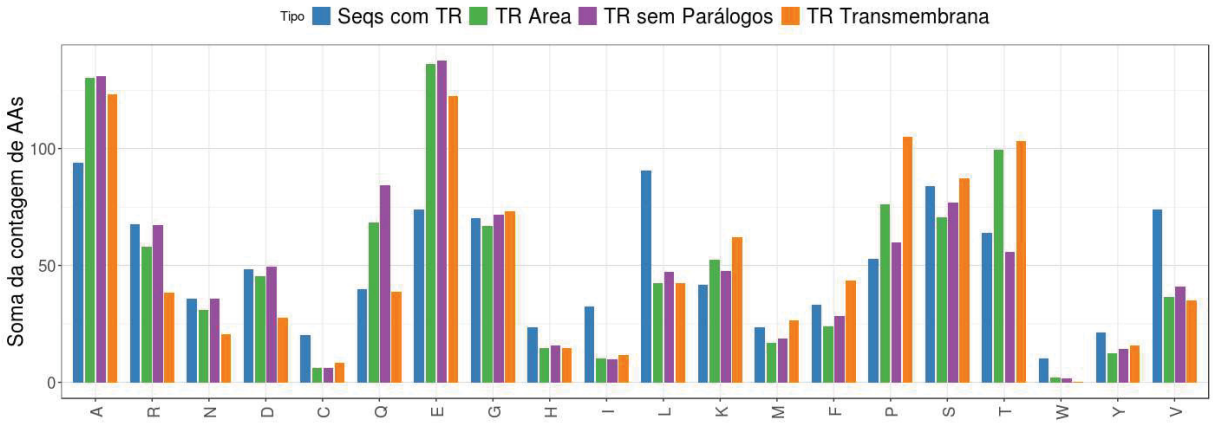
TABELA 3 - PERCENTUAIS DE TRS PARA PROTEÍNAS DE SUPERFÍCIE E GLOBAIS

Tipo	TRs	%	Superfície	%	Demais Seqs	%
Degenerados	1410	65.2%	506	72.7%	904	61.7%
Perfeito	751	34.8%	190	27.3%	561	38.3%
TOTAL	**2161**		**696**		**1465**	

FONTE: A autora (2019).

O GRÁFICO 6 contém as informações de distribuições de AAs para as regiões de TR e para os grupos listados acima: sequências livres de homólogos e genes com domínio transmembrana.

GRÁFICO 6 - COMPARAÇÃO DA CONTAGEM DE AAS NAS REGIÕES DE TR



FONTE: A autora (2019).

NOTA: Gráfico normalizado em frequências por 1.000 AAs para os 3 tipos distintos de áreas de TR. A barra azul, à esquerda, refere-se a contagem de AAs para a sequência completa dos TR e foi mantida somente como referência para comparação ao GRÁFICO 5.

Podemos notar no GRÁFICO 6 que a remoção dos parálogos reduziu a presença de treonina e prolina (P) a níveis normais para todo os genes (barra roxa), um forte indício de que estes sejam AAs comuns a TRs de genes transmembrana. A barra em laranja confirma a maior disponibilidade de prolinas e treoninas em TRs de genes de superfície, acompanhada por uma redução maior em glutamina (Q) e menor em ácido glutâmico (E).

Mendes *et al.* (2013) realizaram uma análise voltada para a presença de aminoácidos potencialmente glicosilados em regiões de TR de genes de proteínas de superfície, já que estas desempenham um papel importante no processo de invasão celular e interação com o sistema humoral do hospedeiro. Asparaginas (N), serinas (S) e treoninas (T) foram o foco da análise. O GRÁFICO 6 confirma a presença de treoninas, um leve aumento com relação a serinas, porém segundo nossa análise asparaginas são menos comuns, mesmo comparadas as sequências de TR completas. É importante ressaltar que no estudo de Mendes et al. (2013) foram consideradas para a análise somente uma lista de TcMUCII para ambos os haplótipos e não um conjunto global de genes com domínios transmembrana, e que sua avaliação considerou a presença dos 3 aminoácidos em conjunto e não individualmente. Além disso, os TRs não selecionados para nosso estudo, mais diversos, podem ser mais ricos em asparaginas do que os presentes em nossa anotação.

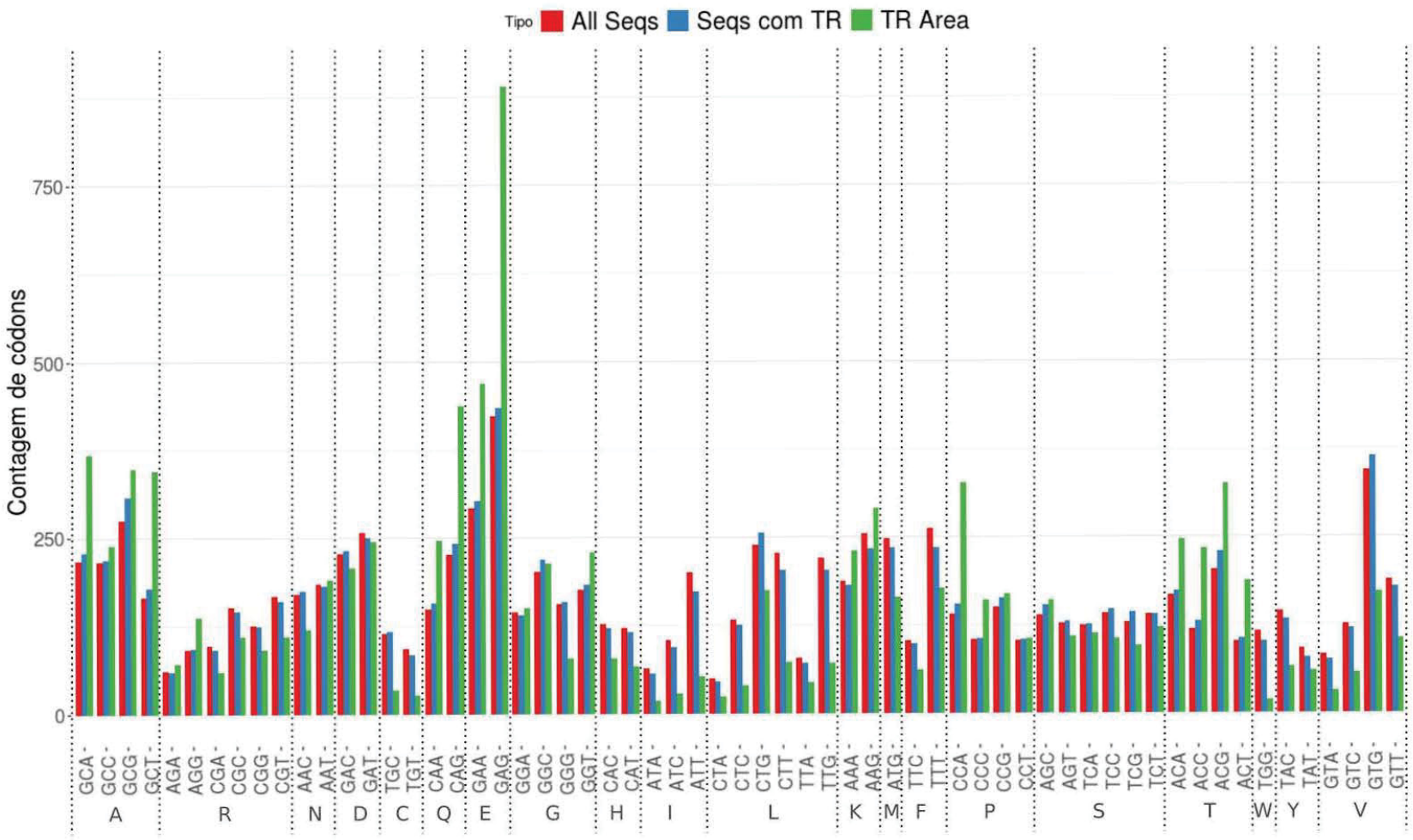
4.1.5 Preferências de códons em região de TR seguem as globais

Após avaliarmos as diferenças em AAs nas regiões de TR, a próxima análise essencial seria a de preferência de códons. Novamente não aplicamos informações de cobertura e efetuamos contagens de códons para as 7.660 sequências, 1.680 sequências com TRs e 2.161 regiões de TRs (GRÁFICO 7). Os códons de parada foram removidos para facilitar a visualização e não acrescentarem informações sobre regiões de TR.

No GRÁFICO 7 podemos observar que os AAs preferenciais nas regiões de TR se destacam, como esperado. Para alanina (A), o códon GCC passa a ser o menos utilizado, enquanto que para os genes completos o GCT é o menos comum. Com relação aos mais utilizados, para os TRs GCA e GCT aparentam ser mais significativos do que o GCG, mais abundante em genes completos.

Já para ácido glutâmico (E), os padrões de códon seguem os das sequências completas, ficando a preferência por GAG ainda mais evidente. A preferência em treonina (T) neste gráfico fica menos evidente devido à distribuição entre 4 códons, porém seus padrões de utilização seguem os dos genes do conjunto completo de sequências. O próximo AA abundante em TRs, glutamina (Q), segue a preferência de códons global, enquanto prolina (P) apresenta uma preferência ainda mais acentuada pelo códon CCA. As reduções esperadas em leucina (L) ficam mais evidentes para os códons CTT, TTG e CTC, enquanto que para valina (V) o códon GTG é o mais afetado. Geramos a aba “C_Counts” no Material suplementar, com os valores de contagens e sua fração relativa a códons, que conferem maior detalhe, em negrito, aos códons preferenciais.

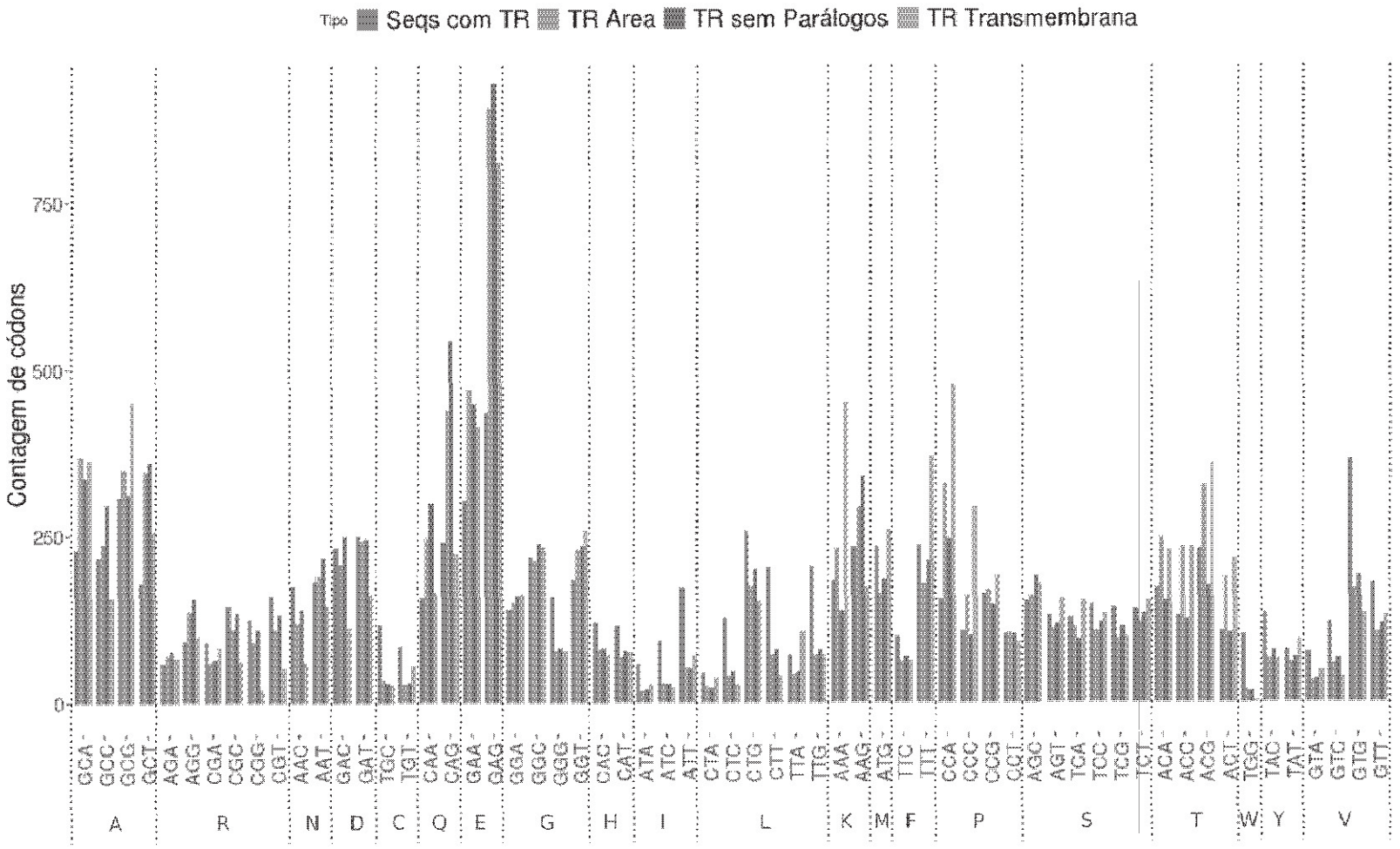
GRÁFICO 7 - PREFERÊNCIA GLOBAL DE CÓDONS



FONTE: A autora (2019).

NOTA: Gráfico de barras normalizado em frequências por 10.000 códons, em ordem alfabética e agrupados por aminoácido.

GRÁFICO 8 - COMPARAÇÃO DA CONTAGEM DE CÓDONS NAS REGIÕES DE TR



FONTE: A autora (2019).

NOTA: Gráfico de barras normalizado em frequências por 10.000 códons, em ordem alfabética e agrupados por aminoácido.

Seguindo com a análise detalhada dos códons de TRs sem a influência de homólogos e isolando somente os genes com domínio transmembrana, geramos o GRÁFICO 8, equivalente ao GRÁFICO 6, para códons.

Para o aminoácido alanina (A) a preferência do códon GCG fica mais evidente para TRs em genes com domínio transmembrana, enquanto que para a relação sem parálogos, a preferência é por GCT. Um dos aminoácidos não tão significativo em região de TR, mas que apresenta uma inversão para genes de superfície em relação aos demais é a lisina (K). A presença de treonina (T) em TRs de genes transmembrana fica novamente evidente, bem como para prolina (P). Adicionamos ao Material Suplementar, aba “C_Counts” os valores de contagens e sua fração relativa, referentes aos 2 grupos adicionais de TRs. Para genes transmembrana verificamos um nível mais alto de divergência entre códons preferenciais com relação ao conjunto sem homólogos e mesmo as todas as sequências. Percebemos, no entanto, que uma parte dessas inversões ocorreu com códons onde a distribuição entre sinônimos já era muito semelhante (cisteína e histidina, por exemplo).

4.1.6 Cobertura de RNA tem pouca influência sobre a preferência de códons

Para avaliarmos se as preferências de códons baseadas em dados do proteoma podem sofrer variações quando expostas a disponibilidade de mRNA, nós aplicamos uma distribuição de cobertura utilizando as leituras de 4 etapas do ciclo de vida disponíveis no transcriptoma sobre a contagem de códons: Tripomastigota TCT, epimastigota, amastigota 24 e 48 hs; os dois últimos representando as fases mais equidistantes dos dois primeiros. Os dados referentes a contagem sem cobertura e às 4 fases descritas acima estão disponíveis no Material Suplementar, aba “C_Counts_Cov”, com códons preferenciais destacados em negrito fase a fase e, para AAs com mais do que 2 sinônimos, os 2 principais foram assinalados. Nesta planilha geramos a proporção de códons para cada aminoácido, o que permite uma percepção mais clara da importância de cada códon a cada etapa do ciclo de vida. Nestes dados podemos verificar algumas mudanças sutis nas preferências:

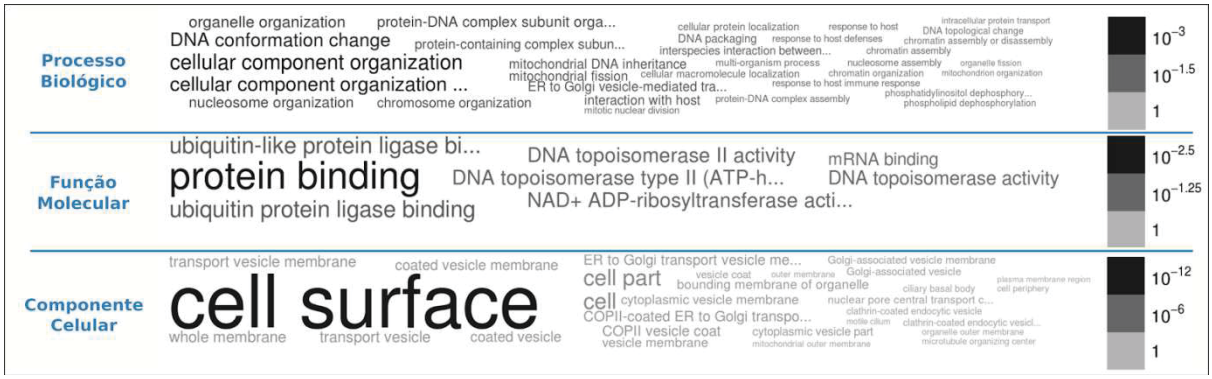
- Alanina (A): teve uma mudança do códon GCA para GCC. Se observarmos os dados sem cobertura (coluna d), seus percentuais eram quase equivalentes;

- Asparagina (N): podemos observar um cenário correspondente ao anterior, com inversão de AAT para AAC;
- Ácido aspártico (D): A mesma situação se repete, com os 2 códons sinônimos estando praticamente balanceados no genoma (GAC 47% e GAT 53%), temos uma inversão com GAC chegando a 54% para epimastigota;
- Serina (S): Em T.cruzi com 6 sinônimos praticamente balanceados, com distribuições no genoma entre 15% e 17% entre todos os códons, verificamos que o códon TCT perde representatividade para o AGC, enquanto AGT e TCA perdem ainda mais importância quando expostos a disponibilidade de mRNA.

4.1.7 Ontologia dos TRs apresenta funções de ligação

Por fim, antes de iniciarmos a análise de epitopos de célula B, avaliamos a função dos genes com TRs, através de pesquisas de GO efetuadas no site TriTrypDB. A FIGURA 9 apresenta a nuvem de palavras referente as 3 ontologias para todo o conjunto de sequências com TRs.

FIGURA 9 - ONTOLOGIAS PARA TODAS AS SEQUÊNCIAS COM TRS.



FONTE: A autora (2019).

Podemos verificar na FIGURA 9 que as proteínas de superfície tem grande influência sobre a ontologia total, possivelmente em decorrência do grande número de homólogos existentes com esta característica. Ao removermos os parálogos verificamos uma distribuição com componentes celulares mais específicos, sendo ainda assim uma boa parte deles relacionadas a componentes de membrana,

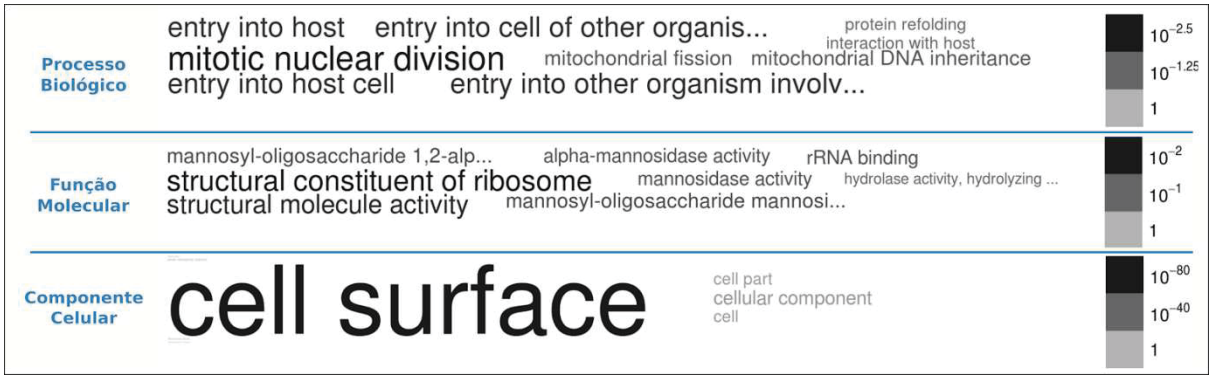
enquanto processos biológicos permanecem com as mesmas características (FIGURA 10).

FIGURA 10 - ONTOLOGIAS PARA SEQUÊNCIAS COM TRS SEM PARÁLOGOS.

FONTE: A autora (2019).

As funções relacionadas a ligações de ubiquitina e topoisomerase perdem um pouco a relevância quando genes parálogos são removidos e outros tipos de ligação ficam mais evidentes. Ao avaliar a ontologia somente dos genes parálogos observamos uma ênfase em processos biológicos relacionados a interações com o hospedeiro e função mais estrutural, o que já era esperado, considerando que quase 50% dos homólogos contém domínio transmembrana (FIGURA 11).

FIGURA 11 - ONTOLOGIAS DOS GENES PARÁLOGOS.

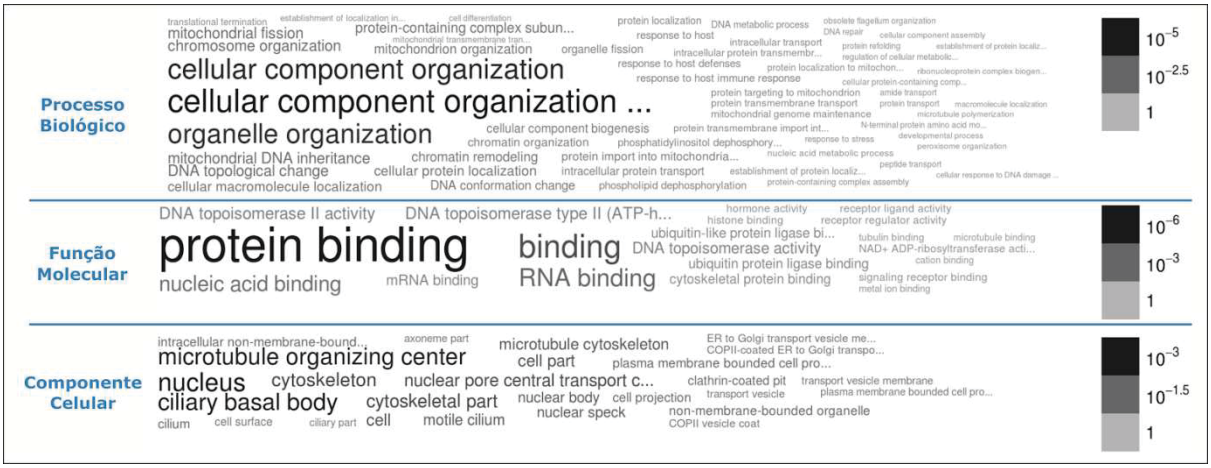


FONTE: A autora (2019).

Ao avaliarmos todo o conjunto dos genes sem a presença de domínio transmembrana (1.151 genes), observamos maior representatividade de

componentes celulares relacionados à citoesqueleto e corpo ciliar, como cinesinas (FIGURA 12).

FIGURA 12 - ONTOLOGIAS DOS GENES COM TR SEM DOMÍNIO TRANSMEMBRANA.



FONTE: A autora (2019).

4.1.8 A detecção de epitopos de célula B é dependente da qualidade das sequências utilizadas

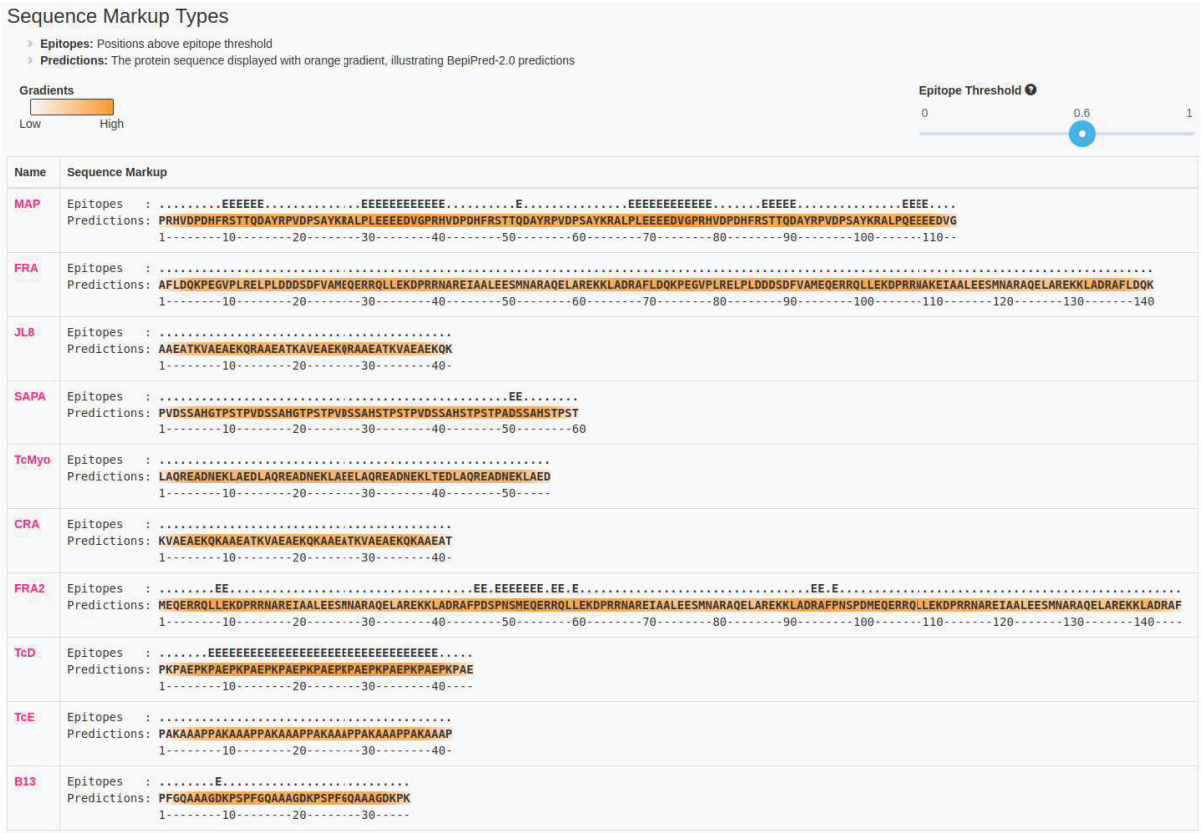
Após a execução desta análise exploratória sobre as regiões de TRs, seguimos com a caracterização de epitopos de célula B. As 1.680 sequências de TR foram submetidas a versão web da ferramenta Bepipred 2.0 em 34 arquivos fasta com até 50 genes, por restrições da ferramenta. Tínhamos a opção de executá-la com somente as regiões de TR, porém para fins comparativos decidimos rodar para os genes completos.

A ferramenta tem como característica retornar probabilidades para cada aminoácido da sequência ao invés da informação de se trata-se ou não de um epitopo (FIGURA 13). Portanto, para auxiliar-nos na definição de uma probabilidade mínima, nós decidimos submeter as regiões de epitopos identificadas em antígenos previamente caracterizados descritas no ANEXO 1 a mesma análise.

O "*slider*" disponível na lateral superior direita da FIGURA 13 permite a seleção da probabilidade mínima de aquele AA ser epitopo e, ao filtrarmos o valor de 0,6, nos deparamos com a marcação de epitopos com região maior do que 7 AAs somente para os peptídeos FRA2, TcD e MAP. Verificamos que ao reduzirmos as probabilidades para 0,54, 51% da área dos peptídeos seriam cobertas e, ao

reduzirmos para 0,5, temos uma cobertura de 76% da área. É importante ressaltar aqui que 0,5 equivale a aleatoriedade, o que nos fez optar pela probabilidade mínima de 0,6.

FIGURA 13 - INTERFACE WEB DA FERRAMENTA BEPIPRED 2.0 PARA EPITOPOS IDENTIFICADOS EM ANTÍGENOS PREVIAMENTE CARACTERIZADOS



FONTE: A autora (2019).

NOTA: Interface gráfica do website da ferramenta Bepipred 2.0 com a detecção dos epitopos identificados em antígenos previamente caracterizados. Consideramos o critério de corte de probabilidade de 0,6, conforme "*slider*" disponível na lateral superior direita da figura, e somente os peptídeos FRA2, TcD e MAP obtiveram cobertura mínima de epitopos, no entanto todos atualmente são utilizados para diagnóstico em *T. cruzi*.

Efetuamos o tratamento do arquivo de saída para nossos 1.680 genes e, além das marcações das regiões de TR, avaliamos também a ocorrência de epitopos para a sequência como um todo que apresentassem probabilidades médias superiores a 0,6 em uma janela móvel de 7 AAs. Adotamos este critério por tratar-se do valor aproximado para regiões de epitopos, conforme a literatura. Verificamos que 90% das sequências apresentavam mais de 1 epitopo, um forte indício de que a

qualidade de predição está totalmente relacionada à qualidade dos dados submetidos à ferramenta, o que reforça a necessidade de elevarmos o critério mínimo para seleção.

Buscamos então filtrar os epitopos com área mínima de 15 AAs dentro das regiões de TR. Apesar de não dispormos de uma medida mínima necessária para avaliação, esta extensão reduziu a ocorrência de TRs muito curtos e simples, mais propensos à reação cruzada com outros organismos.

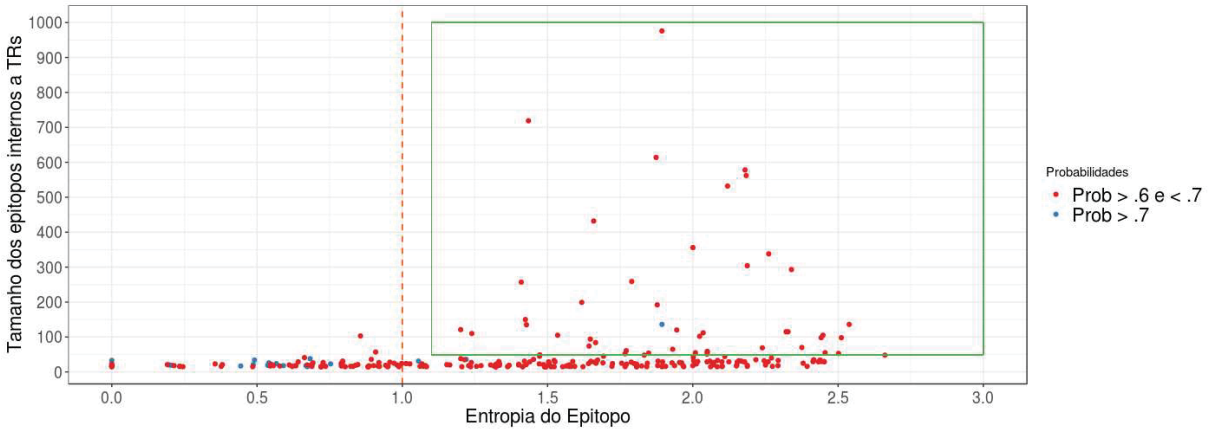
Isolamos então as regiões de TR e consideramos suas médias de probabilidades que atendiam aos critérios mencionados acima. Colunas contendo a região real do epitopo, posição de início, seu tamanho e sua probabilidade foram adicionadas a relação completa de dados dos TRs e disponibilizadas na aba "epitopos_superiores60" do Material Suplementar, totalizando de 273 epitopos internos a 202 regiões de TR.

4.1.9 Epitopos mais prováveis apresentam composição muito simples

Dos 273 epitopos anotados, somente 22 apresentaram uma probabilidade média mínima de 0,7, (10,2%). A probabilidade máxima obtida para 1 AA na região interna aos TRs foi de 0,828, porém nenhuma das anotações teve valores contíguos de 7 AAs maior do 0,8. Os 22 epitopos com probabilidade média de 0,7 apresentam tamanhos entre 15 e 136 AAs e praticamente todos eles são compostos por mono ou di-aminoácidos, com exceção de 4 casos com composição mais dispersa. Regiões de epitopos muito simples apresentam uma tendência em compartilhar sequências conservadas em outros tripanossomatídeos em testes sorológicos, por isso filtramos os TRs com entropia superior a 1,0 utilizando a métrica Entropia de Shannon para as próximas análises deste estudo. O GRÁFICO 9 apresenta a distribuição de tamanhos dos epitopos pela Entropia de Shannon.

Podemos observar que somente 3 casos de epitopos com probabilidade média superior a 0,7 apresentam entropia superior a 1,0, sendo os possíveis melhores candidatos para análises *in vitro*.

GRÁFICO 9 - COMPOSIÇÃO DOS EPÍTOPOS DE CÉLULA B SELECIONADOS



FONTE: A autora (2019).

NOTA: Tamanho dos epitopos selecionados distribuídos de acordo com sua entropia. A linha tracejada vermelha separa 81 epitopos com entropia inferior a 1,0 (30%) e 192 (70%) com entropia superior a 1,0, foco principal das análises futuras deste estudo. A caixa verde delimita os 45 epitopos com tamanho superior a 50 AAs (16,4%).

Para finalizarmos nossa lista de melhores candidatos aos testes de bancada, nós aplicamos as informações de RPKM, das 4 etapas do ciclo de vida selecionadas anteriormente, aos genes nos quais estes epitopos foram anotados. Um alto nível de expressão gênica indicaria maior probabilidade de aquele epitopo ser naturalmente expresso pelo organismo do parasito.

Em razão de não dispormos de uma métrica que delimite qual o valor de expressão que deva ser considerado “alto”, realizamos uma análise nos valores médios de RPKM para as 4 etapas alvo e, como já era esperado, verificamos um acúmulo de valores próximos entre 0 e 150 milhões de *reads* e a existência de *outliers* acima deste valor, com expressão altíssima. Para facilitar nossa compreensão da distribuição esperada, nós geramos os valores limites dos decis, calculados à partir dos desvios da mediana dos valores de RPKM para cada conjunto (QUADRO 4).

QUADRO 4 - LIMITES SUPERIORES DOS DECIS DOS RPKMS POR ETAPA DO CICLO DE VIDA

Etapa / Decil	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
Tripomastigota	19,12	30,01	37,03	44,01	52,05	61,85	75,86	98,49	154,34	8805,54
Epimastigota	5,67	19,52	29,20	37,32	45,85	56,46	72,62	100,78	179,74	12602,10
Amastigota 24hs	5,50	19,63	31,55	40,67	49,61	60,93	75,84	102,75	173,55	13488,94
Amastigota 48hs	5,53	20,43	33,17	42,53	51,42	62,48	77,05	103,99	170,25	11447,29

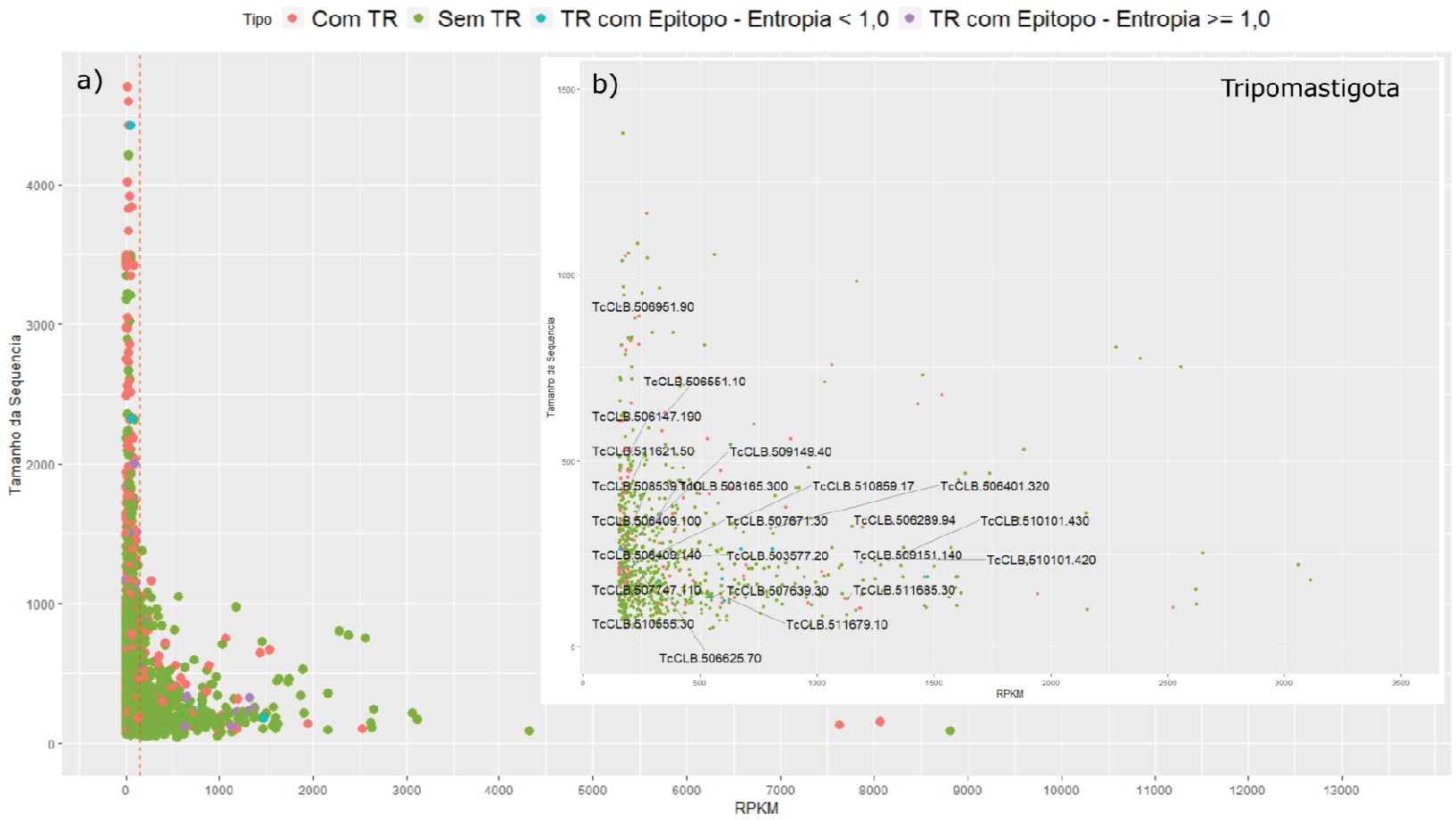
FONTE: A autora (2019).

NOTA: Valores de limites superiores dos decis apresentam certa regularidade para as 4 etapas do ciclo de vida avaliadas. Os limites entre o 9º e 10º decis apresentam uma grande extensão devido a incidência de outliers.

Decidimos não remover da nossa lista os genes com baixos níveis de expressão e sim realizar a ordenação decrescente os registros pelos valores de RPKM para enriquecimento da nossa análise. A aba “epitopos_superiores60” do Material Suplementar recebeu as 4 colunas referentes lista completa de valores de RPKM.

O GRÁFICO 10 exemplifica a distribuição dos genes para a forma tripomastigota para tamanho das sequências relacionado ao seu RPKM. Podemos verificar que, da relação de 273 epitopos internos a TR, somente 24 apresentam bons níveis de expressão e entropia elevada em tripomastigota. Avaliamos então quantos genes apresentam estas características para as 4 etapas avaliadas, e destacamos 2 no Material Suplementar. Entre eles temos alvos já conhecidos (TcCLB.509151.140, TcCLB.506401.320 e TcCLB.509149.40 – proteínas ribossômicas L23a, L7a e L19, respectivamente) e alguns novos alvos que podem ser promissores (TcCLB.510101.430 | ribossômica S21, TcCLB.510859.17 | nucleolar RNA-binding protein, TcCLB.507639.30 | universal minicircle sequence binding protein 1, TcCLB.506625.70 e TcCLB.511621.50| RNA-binding protein, TcCLB.503577.20 | U2 splicing auxiliary factor, TcCLB.507671.30 | 25 kDa translation elongation factor 1-beta, TcCLB.506147.190 | hypothetical protein).

GRÁFICO 10 - DISTRIBUIÇÃO DE RPKM NA FASE TRIPOMASTIGOTA COM DETALHES PARA EPÍTOPOS COM ALTA ENTROPIA



FONTE: A autora (2019).

NOTA: A linha tracejada em vermelho no gráfico a representa o limite do nono decil. O gráfico b apresenta em detalhes a região referente aos 10% dos genes e sequências com TRs com regiões de epitopo e entropia >= 1,0 (cor roxa) tiveram seus Gene IDs impressos.

4.2 ABORDAGEM DE INTELIGÊNCIA ARTIFICIAL

Nesta etapa deste trabalho, nós buscamos aplicar todo o conhecimento adquirido através do processo de anotação e análise de dados no desenvolvimento de uma ferramenta. No entanto, ao invés de utilizarmos os métodos convencionais de desenvolvimento de software, nós optamos pela aplicação de técnicas de inteligência artificial.

4.2.1 Definição de parâmetros do modelo de classificação e atributos de sequências com TRs

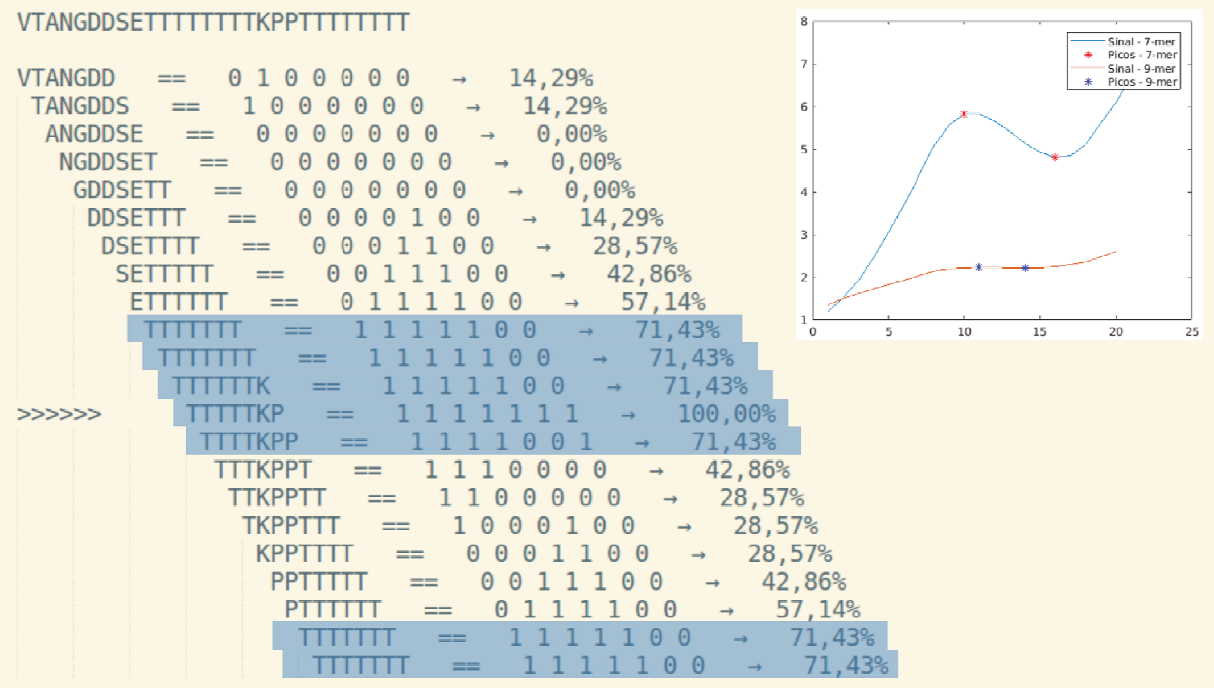
Nós buscamos meios de transformar as sequências de aminoácidos em valores numéricos capazes de representar as características de presença ou ausência de TRs. Foram diversas as métricas geradas com abordagens distintas, com o objetivo de extrair as características mais significativas para diferenciação entre sequências com e sem TRs, atendendo a necessidade do item 2 do processo apresentado na seção 3.2.3 deste trabalho: Engenharia de atributos.

Com o objetivo de garantir que o modelo seja capaz de aprender informações sobre os dados, os atributos selecionados para o classificador utilizam o conceito de janela deslizante de tamanho k (*slide k-mer*) para calcular valores representativos utilizando partes da sequência ao invés de estratifica-la como um todo. Esta técnica mostrou-se mais eficiente na extração de informações representativas do que a alternativa global. A FIGURA 15 exemplifica o processo de *slide* k-mer, dando uma visão para k=7.

As seguintes métricas foram selecionadas para geração do melhor modelo de predição de sequências que apresentam TRs:

a) Diversidade (*divCov*): À partir do número de ocorrências únicas e duplas de aminoácidos no k-mer, estas métricas buscam estimar a diversidade da sequência utilizando probabilidades observadas ou probabilidades *à priori*. Utilizamos a função de Chao (CHAO; CHIU, 2016) uma função interna para geração dos parâmetros (8 parâmetros);

FIGURA 14 - DEMONSTRAÇÃO DO MÉTODO DE EXTRAÇÃO DE MÉTRICAS FUZZY UTILIZANDO TÉCNICA DE SLIDE K-MER



FONTE: A autora (2019).

NOTA: Exemplo do conjunto de métricas fuzzy e o funcionamento do slide k-mer para todos os atributos gerados para o 7-mer "TTTTTKP" comparado fragmento de sequência "VTANGDDSETTTTTTTTKPPTTTTTTTT", extraído dos genes TcCLB.506545.5 e TcCLB.510939.25 (mucina TcMUCI). As linhas destacadas em azul são as efetivamente selecionadas neste exemplo e o gráfico superior à direita demonstra como a contagem de picos foi gerada à partir do sinal representativo da sequência para 7 e 9-mer.

b) Informações estatísticas (*slideStats*): Medidas estatísticas referentes ao mínimo, máximo, média aritmética, desvio padrão e média harmônica da soma das ocorrências únicas de aminoácidos (5 parâmetros);
c) Métricas Fuzzy (*slideFuzzy*): cada k-mer é comparado com a sequência, as ocorrências com no mínimo 70% de identidade são contadas (FIGURA 14) e as métricas de razão da soma das ocorrências superiores a 70% pelo tamanho da sequência, média harmônica das ocorrências e quantidade de picos obtidos à partir de técnica de processamento do sinal de atribuição de pesos de triangularização do k-mer são geradas (3 parâmetros);
d) Características físico-químicas (*phyChem*): Características físico químicas do k-mer foram extraídas e uma média aritmética dos valores foi gerada. Foram consideradas as seguintes características calculadas à partir de funções pertencentes ao pacote de bioinformática do MATLAB: Composição de aminoácidos, Alfa-hélice (Levitt), Folha Beta Antiparalela (Lifson),

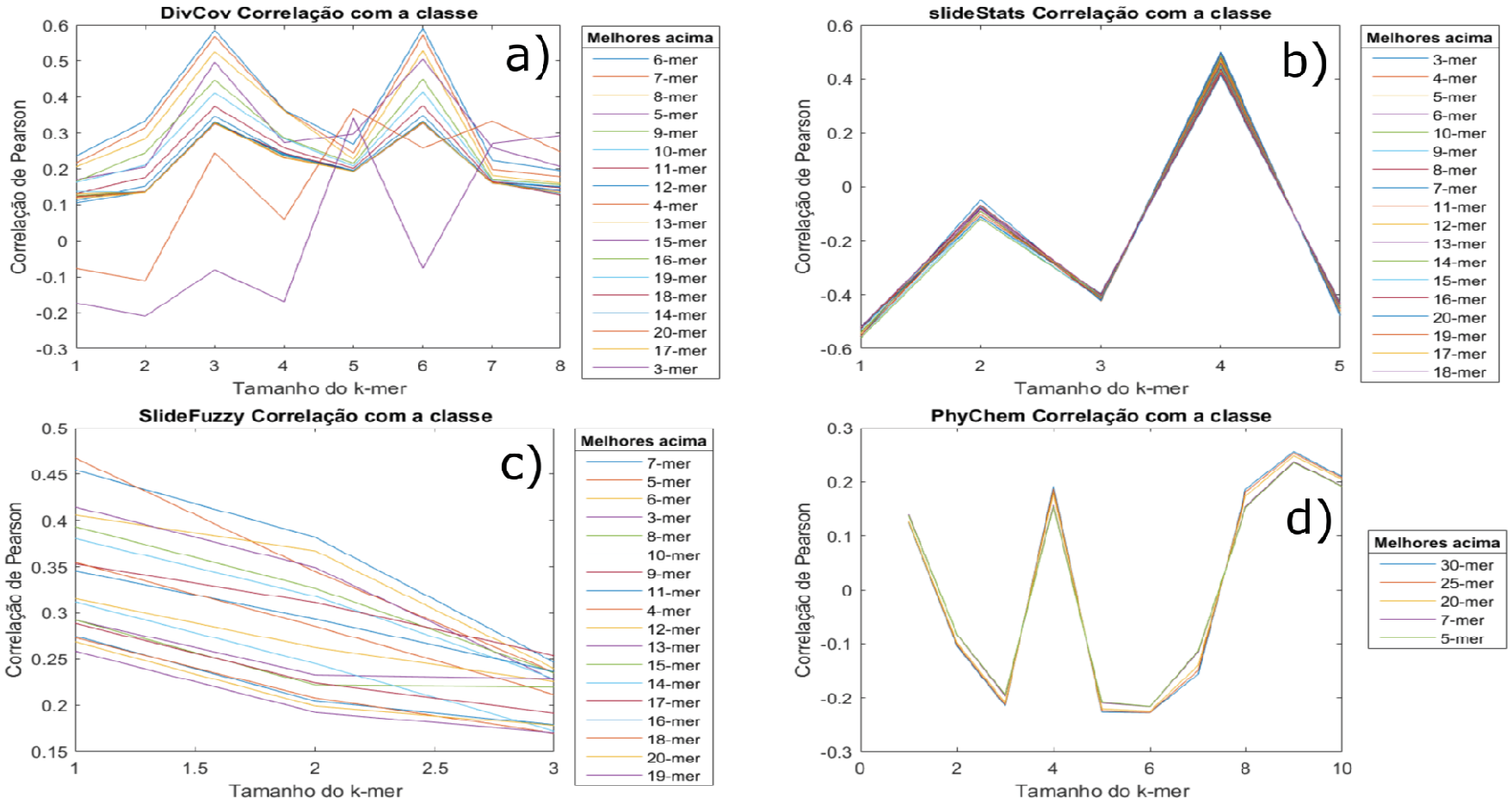
Flexibilidade média (Bhaskaran), Folha Beta (Roux), Bulkiness (Zimmerman), Hidrofobicidade (Kyte&Doolittle), Polaridade(Grantham), Mutabilidade relativa (Dayhoff), Coil (Deleage & Roux) (10 parâmetros);

e) Diferença global normalizada (*normDiff*): Realiza a contagem de uma trinca de aminoácidos com slide 3-mer fixo devido ao custo computacional de memória não justificar o aumento da janela em resultado. A soma das ocorrências é então normalizada pelo tamanho da sequência (1 parâmetro);
f) Combinação de atributos (combFeat): Função que pode ser aplicada a qualquer combinação de atributos, porém não é um atributo em si. Efetua uma transformação baseada em uma função de seno, utilizando a técnica de otimização baseada em IA denominada “algoritmos genéticos”. Esta técnica busca minimizar o erro, neste caso aumentar a correlação, a cada execução através de mudanças sutis na população, mimetizando mutações genéticas. Adotamos uma sistemática de seleção de atributos que contempla os próprios atributos ou os atributos não utilizados na rodada atual de testes (Tamanho variável, dependendo dos atributos de entrada).

Os 6 conjuntos de atributos citados acima foram selecionados através de uma sucessão de testes na geração de modelos. Diversas outras variáveis foram validadas e descartadas por não apresentarem características capazes de gerar informações úteis a serem capturadas por nenhum dos modelos testados.

Além da definição de quais atributos utilizar para avaliação do modelo, foi necessária a definição dos tamanhos de k ideais para os atributos gerados. Para isso nós geramos os valores para k de 3 a 20 para os atributos a (divCov), b (slideStats) e c (slideFuzzy) descritos acima e, devido ao custo computacional, k = (5,7,20,25,30) para o atributo d (phyChem). Antes da seleção dos tamanhos nós efetuamos testes aleatórios para direcionar tamanhos mínimo e máximo com métrica de correlação de Pearson utilizada para comparar os atributos com a classe esperada. O GRÁFICO 11 apresenta as correlações dos 4 conjuntos de atributos ordenada pelos mais correlacionados.

GRÁFICO 11 - CORRELAÇÃO ENTRE ATRIBUTOS DA REDE E ATRIBUTOS



FONTE: A autora(2019).

NOTA: Gráficos apresentam a correlação de cada k-mer com a classe em a) parâmetro divCov, b) parâmetro slideStats, c) parâmetros slideFuzzy e d) parâmetros phyChem. Podemos observar que a variação para slideStats e phyChem é pequena entre os k-mers, indicativo de que o tamanho da janela não é um fator relevante para estas características.

As 3 primeiras métricas apresentaram tendências para k mais curtos, enquanto a 4a teve melhores resultados para valores superiores a 20.

Optamos pela seleção de 5 valores distintos para k para os atributos a, b e c, buscando aumentar a generalização dos dados. Observamos aqui uma tendência a k-mers mais curtos, com valores sempre inferiores a 10. A métrica d obteve melhor valor para 30-mer, por isso seguimos somente com ela em nossas análises.

Uma lista aleatória com os códigos das linhas referentes a cada sequência do conjunto com 7.660 genes foi gerada e particionada em 2 grupos:

- Treinamento: 70% do conjunto, com 5.362 IDs, composta por 1.177 sequências anotadas como Classe 1 - “contém TR” (21,95%);
- Testes: 30% do conjunto, com 2.298 IDs e 503 sequências com Classe 1 (21,89%).

A próxima etapa no processo de geração do preditor é a de geração do modelo (item 3). Seguimos então com os parâmetros padrão conforme descrito na seção 3.2.3 e já iniciamos o item 4, aplicação do modelo em dados de testes e 5, avaliação da performance do modelo. As ações de treinamento e validação do modelo sobre os dados de treinamento, e teste sobre os dados de teste, para as diversas combinações acima tem alto custo computacional. No cenário ideal, todos os tamanhos de k-mer e combinações de atributos deveriam ser testados de forma cruzada para que a definição do melhor custo computacional, de tempo e qualidade dos modelos pudessem ser avaliados, porém esta é uma opção impraticável. Para contornar esta dificuldade algumas definições de execuções mínimas foram necessárias. Neste estudo, nós optamos por utilizar os 5 k-mers mais bem correlacionados com a classe e efetuar a combinação dos atributos, selecionando os 5 valores mais bem classificados para cada gráfico disponível na FIGURA 11. Os 4 primeiros tipos de atributos tiveram a execução adicional individual dos dois k-mers mais bem correlacionados com o objetivo de reduzir tempo de execução, já a redução de atributos equivale aum tempo de geração inferior.

Uma outra característica inerente ao conjunto de dados em estudo foi considerada: o nível de desbalanceamento da amostra. Segundo nossa análise efetuada com a metodologia tradicional, *Tandem Repeats* estão presentes no proteoma do *T. cruzi* em 22% das sequências, o que ocasiona um desbalanceamento de quase 80/20 entre as classes alvo. Apesar de não ser extremo, este cenário já diminui a relevância das características das sequências

com TRs perante o conjunto completo. Isto ocasiona uma tendência de classificações de falsos negativos (FN), isto é, sequências que nós classificamos como “contém TR” são mais difíceis de identificar no conjunto e, consequentemente, classificadas pelos algoritmos como “não contém TR”. Neste ponto precisamos considerar o nosso objetivo e avaliar o quanto a perda de dados da classe 1 impacta nosso processo. Como nosso objetivo é o de utilizar as saídas do 1º preditor para um 2º preditor, o efeito inverso ao mencionado acima pode ser benéfico: aumentar o número de acertos da classe 1 com a penalização da classe 0. Para atingir a este objetivo optamos pelo balanceamento 50/50 das classes no conjunto de treinamento, reduzindo a quantidade de entradas com classe 0 para 1.177 genes aleatórios, gerando o modelo e testando sobre os dados de teste com distribuição real, com os mesmos atributos testados com o conjunto de dados desbalanceados.

Nosso processo de treinamento e validação foi executado, portanto, 114 vezes com diferentes combinações de atributos, entre dados balanceados e desbalanceados e para os 3 algoritmos de ML apontados na seção 3.2.3. Geramos as métricas acurácia, f-1 score, precisão e sensibilidade, porém para avaliarmos a qualidade das predições considerando aumento de acertos na classe 1 e redução na classe 0, somente as métricas sensibilidade e acurácia foram avaliadas. Os valores obtidos, atributos e algoritmos usados foram listados no APÊNDICE 2.

4.2.2 Tempo de execução justifica modelo mais simples com perda moderada

Os valores de sensibilidade obtidos em nossos dados de testes oscilaram de 0,163 a 0,775, enquanto a acurácia variou de 0,604 a 0,877. As piores performances em geral foram obtidas com os parâmetros “phyChem” (Itens 113 e 114 do APÊNDICE 2) e alguns outros atributos individuais, como “slideStats 4-mer” (itens 110 e 117) e “slideFuzzy 7-mer” (item 108), enquanto as melhores variaram entre a combinação “divCov + slideStats + slideFuzzy + normDiff” (item 1), “divCov + slideStats + slideFuzzy + phyChem” (item 13) e “divCov (5,6,7,8,9)-mer” (item 3). Todos os algoritmos de ML apresentaram boa performance para algum dos conjuntos de atributos e, como era esperado, dados balanceados apresentaram em geral melhor sensibilidade. Optamos por adicionalmente considerar a acurácia como critério, pois ela é importante para garantir uma quantidade de erros balanceada na

classe 0 e seu excesso pode representar uma seleção de um conjunto muito grande de sequências com falsos positivos.

Para avaliarmos o real impacto das sequências de TRs mal classificadas na análise, optamos por realizar uma validação dos FN gerados métricas obtinas sobre o conjunto de testes, considerando os seguintes itens do APÊNDICE 2:

a) **Item 1:** Maior sensibilidade, independente da acurácia;
b) **Item 3:** Maior sensibilidade para menor conjunto de atributos;
c) **Item 13:** Acurácia e sensibilidade balanceados para dados balanceados;
d) **Item 40:** Acurácia e sensibilidade balanceados para dados desbalanceados;
e) **Item 69:** Maior acurácia, independente da sensibilidade.

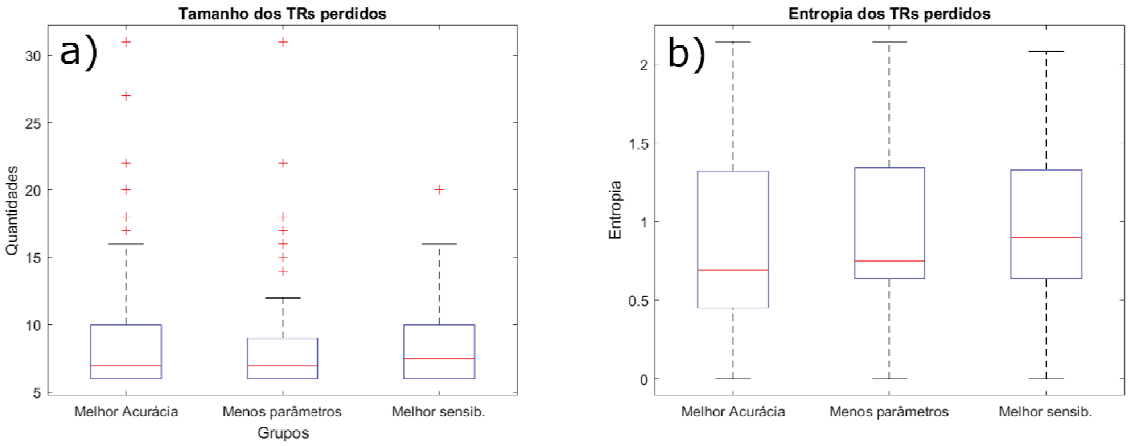
Verificamos que, entre os conjuntos analisados, o que mais apresentou erros de classificação foi o (e), correspondente ao Item 69, com 213 sequências com TR não classificadas. Já o melhor cenário foi o (a), com 113 classificações erradas, seguido de (b), com 114 FNs, ambos modelos gerados sobre sobre um conjunto de treinamento balanceado (FIGURA 15). O GRÁFICO 12 apresenta em mais detalhes as informações de tamanho e entropia dos TRs afetados. Mesmo o pior cenário apontado aqui já apresenta uma boa seleção de TRs, tendo em vista que a próxima rede terá como foco TRs com entropia mais alta e tamanho superior a 15 AAs.

FIGURA 15 - MATRIZES DE CONFUSÃO DOS 5 MODELOS AVALIADOS

	Item 1		Item 3		Item 13		Item 40		Item 69	
Classe Negativa	1536	259	1463	332	1568	227	1624	171	1727	68
Classe Positiva	113	390	114	389	123	380	140	363	213	290
	Predição Negativa	Predição Positiva	Predição Negativa	Predição Positiva	Predição Negativa	Predição Positiva	Predição Negativa	Predição Positiva	Predição Negativa	Predição Positiva

FONTE: A autora (2019).

GRÁFICO 12 - DISTRIBUIÇÃO DOS ERROS DE CLASSIFICAÇÃO DE SEQUÊNCIAS COM TRS.



FONTE: A autora (2019).

NOTA: Box-plot representando as principais características estatísticas dos TRs que não foram selecionados pelos classificadores. Podemos observar em a) que estes são em geral curtos, com tamanhos maiores perdidos na opção “Melhor acurácia” e tamanhos menores em “Melhor sensibilidade”. Em b) verificamos que a entropia média se mantem para todos os grupos próximo a 1,0.

Agora, comparando somente os melhores cenários verificados, precisamos avaliar os fatores quantidade de Falsos Positivos (FP) gerados e velocidade para geração dos parâmetros. Com relação aos FPs, para o melhor modelo temos 259 casos, contra 332 do 2º. O aumento de quase 30% para o último não trará grande impacto de processamento de uma rede secundária. Quanto ao tempo para geração dos parâmetros, temos uma enorme vantagem para o modelo com 2ª melhor performance, com execução em torno de 2 minutos para geração dos 40 atributos contra quase 80 minutos para geração do conjunto completo com 81 colunas. Selecionamos então o modelo baseado nas seguintes configurações: Algoritmo SVM, dados de treinamento balanceados e atributos “divCov (5,6,7,8,9)-mer”. Devido a qualidade obtida no modelo para dados de testes optamos por não efetuar técnicas de validação cruzada adicionais a simples, efetuada sobre dados de treinamento e teste.

5 DISCUSSÃO

Neste trabalho nós realizamos uma análise exploratória de *Tandem Repeats* em *Trypanosoma cruzi*, com enfoque principal em potenciais alvos de anticorpos (presença de epitopo de célula B) para testes sorológicos. A classificação de TRs é um problema complexo, devido sua alta diversidade e possíveis funções desempenhadas. Em *T. cruzi* ainda temos uma enorme carência por uma visão mais ampla de sua composição.

Ao iniciarmos nosso estudo com a escolha das ferramentas para detecção de TRs, fomos capazes de confirmar a afirmação de Pelegrini (2015) de que a utilização de múltiplas ferramentas é imprescindível para aumento na qualidade de anotação. Verificamos que, para o nível de diversidade defendido em nosso estudo, a ferramenta T-REKs mostrou-se uma importante ferramenta de apoio à TRF. Nós confirmamos que a ferramenta TRF, apesar de baseada em anotação de sequências de nucleotídeos, permanece como uma das principais escolhas para TRs menos diversos, justificando o trabalho adicional requerido para adaptação das anotações para sequências de AAs. Em resumo, entendemos que ambas as ferramentas são importantes para a detecção de TRs de baixa diversidade em AAs.

Para selecionarmos os TRs mais adequados, observamos a necessidade de ponderar alguns fatores. Primeiramente, as saídas dos anotadores continham sobreposições, que em nosso estudo consideramos todos os TRs detectados para uma mesma região de sequência, mesmo para casos em que a justaposição referiu-se a apenas um aminoácido. Sua remoção foi necessária para garantir que nossas análises seguintes não fossem distorcidas da informação real obtida da sequência. Outro aspecto importante é o de que nossa busca por antígenos em TRs restringiu nosso escopo de busca aos de menor diversidade. Estes precisam apresentar características de repetição em regiões pequenas, detectáveis por anticorpos, e padrões muito dispersos, compostos por AAs distintos, mesmo que com características físico-químicas semelhantes, ocasionariam a descaracterização da região como repetida. Nosso processo de filtragem requereu um esforço de curadoria manual, pois as técnicas adotadas em estudos anteriores baseavam-se mais na garantia dos TRs mais longos, sem preocupar-se com sua diversidade, ou quando o faziam, era de forma simples, considerando somente a quantidade de GAPs envolvida no alinhamento múltiplo. Percebemos então que a combinação das

métricas diversidade, Jukes-Cantor e SoP em valores avaliados empiricamente seria capaz de atender melhor a todos os pontos destacados acima. É importante ressaltar que os valores de corte foram adaptados para TRs em sequências de *T. cruzi* e que esses parâmetros deverão ser ajustados para utilização em outros organismos. Ao final desta etapa obtivemos o conjunto de TRs que consideramos ideal para nossa análise.

Na avaliação dos TRs anotados foi possível verificarmos uma tendência a regiões repetidas mais simples e curtas, porém TRs longos também foram detectados com sucesso. Nossa técnica de filtragem permitiu a detecção de TRs já conhecidos em *T. cruzi*, como em trans-sialidases (TcCLB.506961.25, TcCLB.507979.30, TcCLB.508607.50, TcCLB.506129.50, TcCLB.508325.230, TcCLB.509265.110, TcCLB.510307.284, TcCLB.510853.40), que já tem sua ocorrência bem documentada (revisado por CARDOSO et al., 2015), proteínas ribossômicas (L19 - TcCLB.509149.40 e TcCLB.509149.60, e L7a - TcCLB.506401.320), tendo as primeiras uma parte da sua região de TR amplamente estudada como epitopos de célula B com boa resposta humoral pelo antígeno TcE (HERNÁNDEZ et al., 2010). Mais de 90% dos TRs apresentam cobertura de até 20% da sua sequência de origem, indicando uma alta representatividade de TRs curtos em *T. cruzi*, considerando-se uma diversidade moderada.

Nós verificamos que a quantidade de TRs degenerados em nossa análise foi apenas 1,8x superior aos perfeitos, ainda menor do que a razão de 3x apontada pelo trabalho de Mendes *et al.* (2013). Esta divergência foi provavelmente ocasionada por diferenças nas sequências e anotações dos TRs, porém não podemos confirmar esta hipótese devido aos dados de TR daquele estudo não terem sido disponibilizados.

Buscamos então avaliar a composição de AAs e códons com relação ao conjunto completo de dados e as sequências onde foram anotados. Esta comparação foi realizada livre de informações de cobertura, pois entendemos que a disponibilidade de mRNA está relacionada à sequência completa e análises sobre partes da sequência poderiam ocasionar distorção no resultado. Ao compararmos a composição de AAs não nos deparamos com nenhuma preferência por classe específica para todo o conjunto de TRs, porém foi possível percebermos que, enquanto a composição das sequências completas que contém TR segue a tendência do conjunto global de sequências (A <-> L > S > V > E >G), as regiões de

TR apresentam preferência de AAs distintas (A > E > T > Q > P), podendo-se destacar de forma mais significativa a alanina (A) e ácido glutâmico (E) e redução de leucina (L), valina (V), triptofano (W), isoleucina (I), cisteína (C) e histidina (H).

Seguimos com a comparação de regiões de TR buscando avaliar AAs para grupos conhecidos. Para isso, buscamos genes com domínio transmembrana, alvos comuns na análise de TRs, e geramos os mesmos dados sem a presença de homólogos. Neste cenário fomos capazes de verificar que o aumento da presença de treoninas (T) e prolinas (P) é ocasionado pelos genes de superfície com TRs, enquanto glutaminas (Q) são mais proeminentes em outros tipos de genes. Dois dos aminoácidos, passíveis de glicolisação, comuns em genes de proteínas de superfície são também comuns em suas regiões de TR (treonina e serina), indicando possível atuação no processo de adesão celular (revisado por BUSCAGLIA et al., 2006). Adicionalmente, avaliamos que 27% das proteínas com domínio transmembrana contem TRs, menos da metade dos quase 60% descritos por Mendes *et al.* (2013) para os 2 haplótipos. Já para o conjunto das proteínas que não são de superfície, Mendes apontou que pouco mais de 20% das proteínas apresentava TRs, enquanto nossas análises contabilizaram pouco mais de 15%. Acreditamos que as divergências avaliadas aqui estão também relacionadas as anotações efetuadas e sequências selecionadas para os estudos. Para TRs degenerados em proteínas de superfície obtivemos 72,7%, bastante superior ao valor obtido por Mendes (~58%), contra 61,7% no restante dos genes com TR, onde o mesmo estudo obteve pouco mais de 20%. Observamos que a quantidade de TRs perfeitos em nosso estudo é representativa, superior a 30%, no entanto está bem distribuída entre os dois grupos. Esta característica acaba contrariando a hipótese de que proteínas de superfície tendem a maior exposição a seleção evolutiva em patógenos extracelulares, apresentando uma tendência a TRs degenerados. Este efeito, no entando, pode ter sido ocasionado por nossa seleção de TRs mais conservados, que removeu uma quantidade considerável de TRs degenerados ou mesmo por características da seleção efetuada por Mendes, dependente de ferramenta única.

Avaliamos também a afirmação de Mendes *et al.* (2013) de que a presença de aminoácidos potencialmente glicosilados (asparagina, treonina e serina) em regiões de TR seria mais abundante em proteínas de superfície do que nas demais, inclusive divergindo dos resultados anteriores de Ramana e Gupta (2009) que indicavam que asparagina teria níveis mais baixos em regiões de TR do que fora

delas. Nós obtivemos os mesmos resultados para treoninas e serinas, confirmando os dois estudos, no entanto os níveis de asparaginas para proteínas de superfície concordam com o estudo de Ramana e Gupta (2009), tendo um nível inferior ao das sequências completas, sendo inclusive mais comuns para TRs em geral do que em TRs transmembrana. Mendes *et al.* (2013) afirmaram que seus níveis de asparaginas eram provavelmente maiores devido a divergências metodológicas, pois suas observações ocorreram predominantemente sobre TRs degenerados, em comparação aos TRs perfeitos avaliados pelo outro estudo. A razão, no entanto, pode estar mais relacionada ao conjunto específico de mucinas utilizado no estudo mais recente, enquanto nosso conjunto de proteínas transmembrana é mais diverso e reforçou os resultados de Ramana.

Em nossas análises de códons nós percebemos variações mais significativas principalmente no conjunto de TRs transmembrana. Para arginina (R) por exemplo, o códon CGA passa a ter o mesmo nível de utilização do que o principal, no entanto, considerando as sequências completas este é o codon menos frequente. Temos uma situação semelhante com histidina (H - CAT), leucina (L - TTA), lisina (K - AAA) e tirosina (Y - TAT), podendo indicar a necessidade de repressão da expressão gênica de algumas das proteínas de superfície. Em uma tentativa de adaptação do CAI utilizando níveis de transcriptoma e proteoma para *T. brucei*, Jeacock *et al.* (2018) obtiveram baixos valores para as glicoproteínas variantes de superfície (VSG), teorizando a mesma necessidade do organismo em reprimir sua expressão. Para o conjunto de TRs sem homólogos as inversão ocorrem com menor frequência e de maneira menos brusca. O mesmo efeito pode ser observado com relação as inversões com a tabela Kazusa, onde os casos de mudanças de preferência são raros.

Buscamos então avaliar a influência da disponibilidade de mRNA no balanço de códons para as diversas etapas do ciclo de vida do *T.cruzi*. Algumas variações podem ser percebidas para Alanina (A), asparagina (N) e ácido aspártico (D) e Serina (S), porém em geral as preferências de códon não variam nem entre as etapas, nem entre elas e o genoma, o que pode ser muito útil em análises de computacionais, pois este é um indicativo de que dados de genoma são capazes de representar a preferência de códons em todas as etapas do ciclo de vida de *T. cruzi*.

É importante ressaltar que a preferência de códons exposta aqui não se aplica a cada gene analisado individualmente, pois sua configuração foi uma das

causas da sua adaptação à dinâmica intracelular a qual ele vem sendo exposto no curso da evolução. A necessidade de aceleração e ou retardamento da síntetise proteica está relacionada com questões que vão desde a conformação tridimensional da proteína até a sinalização do mRNA para degradação, em processos tão complexos que ainda estamos longe de compreendê-los em sua totalidade (MAURO; CHAPPELL, 2014). No entanto, técnicas de otimização de códons vem sendo exploradas e discutidas há muitos anos nas mais diversas áreas de aplicação e atualmente são utilizadas na sintetização de peptídeos antígenos recombinantes, devido ao aumento na velocidade de síntese que ocasionam e, por vezes, garantindo que regiões de difícil replicação em ensaios *in vitro*, ocorram pela simples mudança para um códon mais preferencial (SANTOS et al., 2016).

Em uma análise global das regiões de TR, buscamos informações de ontologia (GO) relacionada a lista geral, lista sem homólogos, somente transmembrana e toda a lista sem genes com domínio transmembrana. Para “processos biológicos” (BP) “organização celular” tem destaque. Para “função molecular” (MF), ligação de membrana é comum a todos os conjuntos, mesmo resultado obtido por Mendes *et al.* (2013). No entanto, podemos perceber variações para “composição celular” (CC), entre superfície celular, componentes de membrana e citoesqueleto. Ao avaliarmos a lista de genes associados aos GOs mencionados, notamos uma grande relação de hipotéticos, resultantes da sua forte presença no proteoma, sem expressão e função conhecidas,o que dificulta o processo de análise de função.

Iniciamos então a etapa de avaliação de TRs buscando sua probabilidade de apresentarem bons candidatos a epitopos de célula B. Ressaltamos aqui que nossa busca foi motivada por antígenos conhecidos como a proteína ribossomica L19 (TcCLB.509149.40 e TcCLB.509149.60) apresentar a composição de TR de baixa diversidade. Em nossa primeira filtragem de regiões de TR com essa característica, utilizamos um critério inicial de corte de 70%, por tratar-se de uma probabilidade razoável, capaz de filtrar bons candidatos.Obtivemos uma lista curta e composta por TRs muito simples, o que motivou a análise dos antígenos mais conhecidos com o intuito de avaliar suas características. Ao avaliarmos o restante dos epitopos para todas as 1.680 sequências, independentemente da sua correlação com regiões de TR, o grande número anotado em uma janela de 7 aminoácidos nos fez questionar o volume de falsos positivos que esta ferramenta realmente gera e, se na submissão

de todos os genes de um organismo, todos apresentarão anotações como ocorrido com as sequências com TRs. Apesar destes questionamentos, estudos de predição *in silico* já foram efetuados em outros patógenos e seus testes em imunoensaio foram capazes de apresentar boa reatividade como antígenos (FARIA et al., 2011), mais um indicativo de que a qualidade das predições está diretamente relacionada à qualidade dos genes apresentados à ferramenta.

Seguimos então com a análise da composição dos TRs selecionados e a simplicidade dos TRs com melhores probabilidade de apresentarem epitopos ocasionou a filtragem pelos de maior entropia interna, pois estes apresentariam menores chances de reatividade cruzada com outros organismos. Decidimos então avaliar quais os melhores candidatos com a perspectiva de disponibilidade de mRNA através dos valores de RPKM. Obtivemos uma lista que contempla alguns antígenos já mapeados e uma relação de 12 novos candidatos com bons níveis de expressão em todas as etapas do ciclo de vida. Se considerarmos os genes mais expressos em tripomastigota nossa lista passa a ser composta por muitas proteínas de superfície, fortemente expressas nesta etapa e, como esperado, suprimidas nas demais. Esta análise específica pode permitir um aumento significativo nos possíveis alvos, direcionados para etapas do ciclo de vida onde são mais expressos.

Ao final de todo o processo de análise iniciamos o processo de geração um bom preditor de sequências que contem ou não TRs de diversidade moderada, característica que tem se mostrado relevante para os alvos de sorologia estudados até então. A construção desta ferramenta foi motivada pelas dificuldades encontradas no processo de análise, que iniciaram-se com a série de ferramentas de TR disponíveis, seguida da tarefa de filtragem dos candidatos mais adequados e por fim das limitações encontradas para anotação de epitopos de célula B, limitada a poucas sequências por vez. Neste processo, nós percebemos que alguns padrões poderiam ser capturados por um modelo de IA robusto.

Tivemos uma grande dificuldade na definição de parâmetros capazes apresentar bons padrões. Boa parte das variáveis testadas tem a característica de diferenciar sequências pelo acúmulo de baixa diversidade. Ao tentarmos expô-los a anotações globais de TR, sem o filtro efetuado por nós para remoção dos mais dispersos, a qualidade das predições foi drasticamente reduzida. Apesar de este ser um aspecto limitante para uma ferramenta global de anotação de TRs, apresentou-se como uma vantagem para nosso objetivo de busca de alvos de sorologia, tendo

em vista que os antígenos mais comumente avaliados de *T. cruzi* apresentam esta característica de TRs mais conservados.

Após diversas tentativas de transformação com o intuito de gerarmos extratificações matemáticas das sequências, geramos alguns conjuntos de características que apresentaram maior sucesso na classificação dos TRs com o nível de diversidade obtido na análise de dados efetuada anteriormente neste estudo. O aumento nos acertos de classificação de sequências com TRs foi gradativo e, em razão disso, buscamos catalogar um conjunto grande de características capazes de melhorar a performance dos algoritmos implementados.

Em uma avaliação global dos atributos testados, observamos que características físico-químicas aparentam ter pouca influência de forma independente sobre este tipo de TR, dado que seus percentuais de acurácia ficaram entre os piores para alguns modelos (60 e 65%) e, mesmo para os casos de acurácias melhores, apresentaram sensibilidades muito ruins (16 a 18% para alguns casos), indicando uma baixa capacidade de diferenciação de padrões entre regiões com e sem TRs. Nossa análise realmente demonstrou que as preferências de aminácidos não referem-se a tipos específicos, o que corrobora com os resultados obtidos pelo preditor. Em geral a combinação de atributos com múltiplos k-mers apresentou melhores resultados do que os individuais. No entanto, ao buscarmos os k-mers que apresentassem melhores correlações com as classes, podemos perceber que uma preferência para valores entre 5 e 10 AAs, o que pode indicar que as transformações realizadas podem estar capturando justamente características de conformação proteica, como folhas-beta e alfa-hélices.

O modelo selecionado ao final do processo foi capaz de obter uma acurácia de 80% e sensibilidade de 77% e verificamos que os *Tandem Repeats* mais curtos e simples compuseram os principais casos de erros, possivelmente devido sua baixa representatividade na sequência como um todo. A quantidade de acertos e erros do modelo varia de acordo com as necessidades do processo e a decisão de quais métricas devem ser utilizadas para sua validação dependem da avaliação dessa necessidade (FLACH, 2012). Ao verificarmos que nosso preditor era capaz avaliar a existência de TRs de baixa diversidade em uma sequência e que esta diversidade era equivalente a apresentada por muitos dos antígenos de *T. cruzi* , percebemos uma oportunidade de anotação de sequências com escopo de antígenos. Definimos então que a tolerância a erros poderia ser mais alta e que falsos positivos poderiam

ser filtrados na etapa de anotação de epitopos de células B. Portanto, as métricas acurácia e sensibilidade nos valores apresentados pelo modelo escolhido seriam suficientes para atender a necessidade avaliada. No entanto, sua escolha não deveu-se somente aos valores obtidos nas métricas de seleção, mas também ao tempo de geração de suas variáveis ser muito inferior ao requerido pelas demais avaliadas. Entendemos que o maior número de acertos obtido com todo o conjunto de métricas não justifica o tempo de execução tão superior demandado pela geração das características.

Ao avaliarmos as variáveis selecionadas por este estudo, o conjunto o qual nós denominamos “DivCov” (diversidade de cobertura), sua capacidade de caracerização de quão diversa é a sequência foi mais eficiente do que as técnicas utilizadas pelas demais variáveis para atender este mesmo objetivo. Uma de suas métricas, o estimador de riqueza de Chao, muito utilizada na análise comparativa de biodiversidade de espécies em estudos de ecologia, mostrou-se uma métrica eficiente na avaliação de diversidade intra sequência, em conjunto com a métrica desenvolvida neste trabalho, riqueza de Raittz. Uma análise mais aprofundada faz-se necessária para exploração de todo o seu potencial na avaliação de nível de diversidade de sequências.

Nosso preditor, portanto, é capaz de atender a necessidade delimitada de garantir que os TRs suficientemente diversos sejam corretamente anotados, reduzindo a quantidade de sequências a serem submetidas a uma ferramenta de predição de epitopo de célula B conforme esperado.

Entendemos que os objetivos específicos deste trabalho foram atendidos através da caracterização de TRs em *T. cruzi*, seleção de ferramentas e técnicas adequadas para sua identificação, avaliação de cobertura utilizando dados de mRNA e anotação de epitopos de células B nessas regiões. Tivemos sucesso na geração um modelo de inteligência artificial capaz de extrair características significativas de sequências que contém TRs. Este foi o primeiro passo para geração de uma ferramenta robusta, que tornará o processo de identificação de TRs com diversidade propícia para regiões de antígenos de patógenos eucariotos mais eficaz e eficiente.

6 CONCLUSÃO

Neste trabalho nós caracterizamos um pouco mais *Tandem Repeats* em *Trypanosoma cruzi*, de forma mais abrangente do que estudos anteriores. Verificamos alguns dos padrões de preferências de aminoácidos dessas regiões, reforçando algumas informações já obtidas para proteínas de superfície e enriquecendo o conhecimento de TRs em geral. Geramos uma nova tabela de preferências de códons que corrobora em grande parte com a tabela mais utilizada, mas também geramos para regiões de TR, que podem ser exploradas para obtermos mais conhecimento sobre como os códons sinônimos podem influenciar nos níveis de produção proteica e auxiliar no processo de síntese e validação de epitopos detectáveis em testes sorológicos. Através da anotação de epitopos de célula B em regiões de TR, geramos uma relação mais ampla e direcionada de candidatos a testes diagnóstico. À partir do aprendizado obtido com todo este processo fomos capazes de gerar um preditor de sequências de TR de baixa diversidade, sendo este o primeiro passo para desenvolvimento de um processo mais simples e prático para detecção de TRs epitopos de células B, não somente para *T. cruzi*, mas para outros patógenos humanos.

6.1 RECOMENDAÇÕES PARA TRABALHOS FUTUROS

Podemos apontar como possíveis trabalhos futuros:

- Validação do nível de homologia dos epitopos e TRs mapeados com sequências de outros organismos;
- Criação de um novo índice capaz de classificar epitopos de célula B mais adaptados aos códons preferenciais;
- Experimento *in vitro* utilizando os melhores epitopos de célula B anotados por este trabalho utilizando a melhor configuração de códons.
- Expansão do modelo para predição de presença e ausência de TRs na sequência, buscando sua localização e aplicação em outros organismos patógenos eucariotos;
- Validação de performance do preditor de TR utilizando redes neurais convolucionais (RNC), validações cruzadas para melhoria de

parâmetros e desenvolvimento de nova versão utilizando a linguagem Python;

- Encaminhamento do modelo de predição de região de epitopo de célula B já desenvolvido e com testes preliminares positivos, para geração de uma ferramenta que anote TRs e epitopos de células B automaticamente.

REFERÊNCIAS

AKASHI, H.; EYRE-WALKER, A. Translational selection and molecular evolution. **Current opinion in genetics & development**, v. 8, n. 6, p. 688–693, 1998.

ANDRADE, M. A.; PEREZ-IRATXETA, C.; PONTING, C. P. Protein Repeats: Structures, Functions, and Evolution. **Journal of Structural Biology**, 2001. Disponível em: <http://dx.doi.org/10.1006/jsbi.2001.4392>. .

ANSARI, H.; RAGHAVA, G. P. S. Identification of conformational B-cell Epitopes in an antigen from its primary sequence. **Immunome Research**, 2010. Disponível em: <http://dx.doi.org/10.1186/1745-7580-6-6>. .

DE AVALOS, S. V.; BLADER, I. J.; FISHER, M.; BOOTHROYD, J. C.; BURLEIGH, B. A. Immediate/Early Response toTrypanosoma cruziInfection Involves Minimal Modulation of Host Cell Transcription. **The Journal of biological chemistry**, v. 277, n. 1, p. 639–644, 2001.

BALOUZ, V.; AGÜERO, F.; BUSCAGLIA, C. A. Chagas Disease Diagnostic Applications: Present Knowledge and Future Steps. **Advances in parasitology**, v. 97, p. 1–45, 2017.

BAPTISTA, R. P.; REIS-CUNHA, J. L.; DEBARRY, J. D.; et al. Assembly of highly repetitive genomes using short reads: the genome of discrete typing unit III Trypanosoma cruzi strain 231. **Microbial genomics**, 2018. Disponível em: <http://dx.doi.org/10.1099/mgen.0.000156>. .

BARLOW, D. J.; EDWARDS, M. S.; THORNTON, J. M. Continuous and discontinuous protein antigenic determinants. **Nature**, 1986. Disponível em: <http://dx.doi.org/10.1038/322747a0>. .

BENSON, G. Tandem repeats finder: a program to analyze DNA sequences. **Nucleic Acids Research**, 1999. Disponível em: <http://dx.doi.org/10.1093/nar/27.2.573>. .

BERNÁ, L.; RODRIGUEZ, M.; CHIRIBAO, M. L.; et al. Expanding an expanded genome: long-read sequencing of Trypanosoma cruzi. **Microbial genomics**, v. 4, n. 5, 2018. Disponível em: <http://dx.doi.org/10.1099/mgen.0.000177>. .

BERN, C. Chagas' Disease. **The New England journal of medicine**, v. 373, n. 5, p. 456–466, 2015.

BIEGERT, A.; SÖDING, J. De novo identification of highly diverged protein repeats by probabilistic consistency. **Bioinformatics** , v. 24, n. 6, p. 807–814, 2008.

BLYTHE, M. J.; FLOWER, D. R. Benchmarking B cell epitope prediction: underperformance of existing methods. **Protein science: a publication of the Protein Society**, v. 14, n. 1, p. 246–248, 2005.

BUSCAGLIA, C. A.; CAMPO, V. A.; FRASCH, A. C. C.; DI NOIA, J. M. Trypanosoma cruzi surface mucins: host-dependent coat diversity. **Nature reviews**. Microbiology, v. 4, n. 3, p. 229–236, 2006.

CAPUANI, L.; BIERRENBACH, A. L.; PEREIRA ALENCAR, A.; et al. Mortality among blood donors seropositive and seronegative for Chagas disease (1996-2000) in São Paulo, Brazil: A death certificate linkage study. **PLoS neglected tropical diseases**, v. 11, n. 5, p. e0005542, 2017.

CARDOSO, M. S.; REIS-CUNHA, J. L.; BARTHOLOMEU, D. C. Evasion of the Immune Response by Trypanosoma cruzi during Acute Infection. **Frontiers in immunology**, v. 6, p. 659, 2015.

CHAO, A.; CHIU, C.-H. Nonparametric Estimation and Comparison of Species Richness. **eLS**, 2016. Disponível em: <http://dx.doi.org/10.1002/9780470015902.a0026329>. .

CLAYTON, C. E. Gene expression in Kinetoplastids. **Current opinion in microbiology**, v. 32, p. 46–51, 2016.

DELGRANGE, O.; RIVALS, E. STAR: an algorithm to Search for Tandem Approximate Repeats. **Bioinformatics** , v. 20, n. 16, p. 2812–2820, 2004.

DIAS, J. C. P.; RAMOS, A. N., Jr; GONTIJO, E. D.; et al. 2 nd Brazilian Consensus on Chagas Disease, 2015. **Revista da Sociedade Brasileira de Medicina Tropical**, v. 49Suppl 1, n. Suppl 1, p. 3–60, 2016.

DUTRA, W. O.; MENEZES, C. A. S.; MAGALHÃES, L. M. D.; GOLLOB, K. J. Immunoregulatory networks in human Chagas disease. **Parasite immunology**, v. 36, n. 8, p. 377–387, 2014.

EL-SAYED, N. M. The Genome Sequence of Trypanosoma cruzi, Etiologic Agent of Chagas Disease. **Science**, v. 309, n. 5733, p. 409–415, 2005.

EPTING, C. L.; COATES, B. M.; ENGMAN, D. M. Molecular mechanisms of host cell invasion by Trypanosoma cruzi. **Experimental parasitology**, v. 126, n. 3, p. 283–291, 2010.

FARIA, A. R.; COSTA, M. M.; GIUSTA, M. S.; et al. High-throughput analysis of synthetic peptides for the immunodiagnosis of canine visceral leishmaniasis. **PLoS neglected tropical diseases**, v. 5, n. 9, p. e1310, 2011.

Fiocruz.**Tabela de códons gerada pelo Instituto Kazusa** - Disponível em: <http://www.dbbm.fiocruz.br/TcruziDB/codon.html> Acesso em: 05 mai. 2019

FLACH, P. Machine Learning: The Art and Science of Algorithms that Make Sense of Data. **Cambridge University Press**, 2012.

FRANZÉN, O.; OCHAYA, S.; SHERWOOD, E.; et al. Shotgun Sequencing Analysis of Trypanosoma cruzi I Sylvio X10/1 and Comparison with *T. cruzi* VI CL Brener. **PLoS neglected tropical diseases**, v. 5, n. 3, p. e984, 2011.

GASCON, J.; BERN, C.; PINAZO, M.-J. Chagas disease in Spain, the United States and other non-endemic countries. **Acta tropica**, v. 115, n. 1-2, p. 22–27, 2010.

GOMES, Y. M.; LORENA, V. M. B.; LUQUETTI, A. O. Diagnosis of Chagas disease: what has been achieved? What remains to be done with regard to diagnosis and follow up studies? **Memorias do Instituto Oswaldo Cruz**, v. 104 Suppl 1, p. 115–121, 2009.

GUIZETTI, J.; SCHERF, A. Silence, activate, poise and switch! Mechanisms of antigenic variation in Plasmodium falciparum. **Cellular microbiology**, v. 15, n. 5, p. 718–726, 2013.

GÜRTLER, R. E.; CARDINAL, M. V. Reservoir host competence and the role of domestic and commensal hosts in the transmission of Trypanosoma cruzi. **Acta tropica**, v. 151, p. 32–50, 2015.

HEGER, A.; HOLM, L. Rapid automatic detection and alignment of repeats in protein sequences. **Proteins**, v. 41, n. 2, p. 224–237, 2000.

HERNÁNDEZ, P.; HEIMANN, M.; RIERA, C.; et al. Highly effective serodiagnosis for Chagas' disease. **Clinical and vaccine immunology**: CVI, v. 17, n. 10, p. 1598–1604, 2010.

HOPP, T. P.; WOODS, K. R. Prediction of protein antigenic determinants from amino acid sequences. **Proceedings of the National Academy of Sciences**, 1981. Disponível em: <http://dx.doi.org/10.1073/pnas.78.6.3824>. .

HORN, D. Codon usage suggests that translational selection has a major impact on protein expression in trypanosomatids. **BMC genomics**, v. 9, n. 1, p. 2, 2008.

JACKSON, Y.; PINTO, A.; PETT, S. Chagas disease in Australia and New Zealand: risks and needs for public health interventions. **Tropical medicine & international health: TM & IH**, v. 19, n. 2, p. 212–218, 2014.

JEACOCK, L.; FARIA, J.; HORN, D. Codon usage bias controls mRNA and protein abundance in trypanosomatids. **eLife**, v. 7, 2018. Disponível em: <http://dx.doi.org/10.7554/eLife.32496>. .

JESPERSEN, M. C.; PETERS, B.; NIELSEN, M.; MARCATILI, P. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. **Nucleic acids research**, v. 45, n. W1, p. W24–W29, 2017.

JORDA, J.; KAJAVA, A. V. T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm. **Bioinformatics** , v. 25, n. 20, p. 2632–2638, 2009.

KAJAVA, A. V. Tandem repeats in proteins: from sequence to structure. **Journal of structural biology**, v. 179, n. 3, p. 279–288, 2012.

KOOHY, H. The rise and fall of machine learning methods in biomedical research. **F1000Research**, v. 6, p. 2012, 2017.
LERAT, E. Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. **Heredity**, v. 104, n. 6, p. 520–533, 2010.

LIBBRECHT, M. W.; NOBLE, W. S. Machine learning applications in genetics and genomics. **Nature Reviews Genetics**, 2015. Disponível em: <http://dx.doi.org/10.1038/nrg3920>. .

LI, Y.; SHAH-SIMPSON, S.; OKRAH, K.; et al. Transcriptome Remodeling in Trypanosoma cruzi and Human Cells during Intracellular Infection. **PLoS pathogens**, v. 12, n. 4, p. e1005511, 2016.

LOPES, R. DA S.; MORAES, W. J. L.; RODRIGUES, T. DE S.; BARTHOLOMEU, D. C. ProGeRF: proteome and genome repeat finder utilizing a fast parallel hash function. **BioMed research international**, v. 2015, p. 394157, 2015.

MAIR, G.; SHI, H.; LI, H.; et al. A new twist in trypanosome RNA metabolism: cis-splicing of pre-mRNA. **RNA** , v. 6, n. 2, p. 163–169, 2000.

MARTINS-MELO, F. R.; RAMOS, A. N., Jr; ALENCAR, C. H.; HEUKELBACH, J. Mortality due to Chagas disease in Brazil from 1979 to 2009: trends and regional differences. **Journal of infection in developing countries**, v. 6, n. 11, p. 817–824, 2012.

MATLAB **Documentation website** - Disponível em: *<https://www.mathworks.com/help/matlab/live-scripts-and-functions.html>* Acesso em: 06 ago. 2018

MAURO, V. P.; CHAPPELL, S. A. A critical analysis of codon optimization in human therapeutics. **Trends in molecular medicine**, v. 20, n. 11, p. 604–613, 2014.

MENDES, T. A. O.; LOBO, F. P.; RODRIGUES, T. S.; et al. Repeat-enriched proteins are related to host cell invasion and immune evasion in parasitic protozoa. **Molecular biology and evolution**, v. 30, n. 4, p. 951–963, 2013.

MONCAYO, A.; SILVEIRA, A. C. Current epidemiological trends for Chagas disease in Latin America and future challenges in epidemiology, surveillance and health policy. **Memorias do Instituto Oswaldo Cruz**, v. 104 Suppl 1, p. 17–30, 2009.

MÜLLER, A. C.; GUIDO, S. Introduction to Machine Learning with Python: A Guide for Data Scientists. **O'Reilly Media, Inc**., 2016.

NAGARKATTI, R.; BIST, V.; SUN, S.; et al. Development of an aptamer-based concentration method for the detection of Trypanosoma cruzi in blood. **PloS one**, v. 7, n. 8, p. e43533, 2012.

NEWMAN, A. M.; COOPER, J. B. XSTREAM: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. **BMC bioinformatics**, v. 8, p. 382, 2007.

DE LIMA NICHIO, B. T.; DE OLIVEIRA, A. M. R.; DE PIERRI, C. R.; et al. RAFTS3G - An efficient and versatile clustering software to analyses in large protein datasets. **bioRxiv**, 3. set. 2018. Disponível em: <https://www.biorxiv.org/content/10.1101/407437v1.abstract>. Acesso em: 21 nov. 2018.

PADILLA, A. M.; SIMPSON, L. J.; TARLETON, R. L. Insufficient TLR activation contributes to the slow development of CD8+ T cell responses in Trypanosoma cruzi infection. **Journal of immunology** , v. 183, n. 2, p. 1245–1252, 2009.

PARHAM, P. The Immune System, 3rd Edition. **Garland Science**, 2009.

PELLEGRINI, M. Tandem Repeats in Proteins: Prediction Algorithms and Biological Role. **Frontiers in bioengineering and biotechnology**, v. 3, p. 143, 2015.

PELLEQUER, J. L.; WESTHOF, E.; VAN REGENMORTEL, M. H. Predicting location of continuous epitopes in proteins from their primary structures. **Methods in enzymology**, v. 203, p. 176–201, 1991.

PELLEQUER, J. L.; WESTHOF, E.; VAN REGENMORTEL, M. H. Correlation between the location of antigenic sites and the prediction of turns in proteins. **Immunology letters**, v. 36, n. 1, p. 83–99, 1993.

PETERSEN, B.; PETERSEN, T. N.; ANDERSEN, P.; NIELSEN, M.; LUNDEGAARD, C. A generic method for assignment of reliability scores applied to solvent accessibility predictions. **BMC structural biology**, v. 9, p. 51, 2009.

PIACENZA, L.; PELUFFO, G.; ALVAREZ, M. N.; MARTÍNEZ, A.; RADI, R. Trypanosoma cruzi antioxidant enzymes as virulence factors in Chagas disease. **Antioxidants & redox signaling**, v. 19, n. 7, p. 723–734, 2013.

PIOVESAN, A.; VITALE, L.; PELLERI, M. C.; STRIPPOLI, P. Universal tight correlation of codon bias and pool of RNA codons (codonome): The genome is optimized to allow any distribution of gene expression values in the transcriptome from bacteria to humans. **Genomics**, v. 101, n. 5, p. 282–289, 2013.

PITCOVSKY, T. A.; BUSCAGLIA, C. A.; MUCCI, J.; CAMPETELLA, O. A Functional Network of Intramolecular Cross-Reacting Epitopes Delays the Elicitation of Neutralizing Antibodies toTrypanosoma cruzi trans-Sialidase. **The Journal of infectious diseases**, v. 186, n. 3, p. 397–404, 2002.

RAMANA, J.; GUPTA, D. ProtVirDB: a database of protozoan virulent proteins. **Bioinformatics** , v. 25, n. 12, p. 1568–1569, 2009.

RAMÍREZ, J. C.; CURA, C. I.; DA CRUZ MOREIRA, O.; et al. Analytical Validation of Quantitative Real-Time PCR Methods for Quantification of Trypanosoma cruzi DNA in Blood Samples from Chagas Disease Patients. **The Journal of molecular diagnostics: JMD**, v. 17, n. 5, p. 605–615, 2015.

RASSI, A.; RASSI, A.; MARIN-NETO, J. A. Chagas disease. **The Lancet**, v. 375, n. 9723, p. 1388–1402, 2010.

REINA-SAN-MARTIN, B.; COSSON, A.; MINOPRIO, P. Lymphocyte polyclonal activation: a pitfall for vaccine design against infectious agents. **Parasitology today** , v. 16, n. 2, p. 62–67, 2000.

DOS REIS, M.; WERNISCH, L. Estimating translational selection in eukaryotic genomes. **Molecular biology and evolution**, v. 26, n. 2, p. 451–461, 2009.

RICHARD, F. D.; ALVES, R.; KAJAVA, A. V. Tally: a scoring tool for boundary determination between repetitive and non-repetitive protein sequences. **Bioinformatics**, 2016. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/btw118>. .

ROTH, A.; ANISIMOVA, M.; CANNAROZZI, G. M. Measuring codon usage bias. **Codon Evolution**. p.189–217, 2012.

SALIMI, N.; FLERI, W.; PETERS, B.; SETTE, A. Design and utilization of epitope-based databases and predictive tools. **Immunogenetics**, v. 62, n. 4, p. 185–196, 2010.

SANTOS, F. L. N.; CELEDON, P. A. F.; ZANCHIN, N. I. T.; et al. Performance Assessment of Four Chimeric Trypanosoma cruzi Antigens Based on Antigen-Antibody Detection for Diagnosis of Chronic Chagas Disease. **PloS one**, v. 11, n. 8, p. e0161100, 2016.

SCHAPER, E.; KAJAVA, A. V.; HAUSER, A.; ANISIMOVA, M. Repeat or not repeat?—Statistical validation of tandem repeat prediction in genomic sequences. **Nucleic Acids Research**, 2012. Disponível em: <http://dx.doi.org/10.1093/nar/gks726>. .

SCHAPER, E.; KORSUNSKY, A.; PEČERSKA, J.; et al. TRAL: tandem repeat annotation library. **Bioinformatics** , v. 31, n. 18, p. 3051–3053, 2015.

SEGOVIA, M.; CARRASCO, H. J.; MARTÍNEZ, C. E.; et al. Molecular epidemiologic source tracking of orally transmitted Chagas disease, Venezuela**. Emerging infectious diseases**, v. 19, n. 7, p. 1098–1101, 2013.

SIMPSON, L. Kinetoplast DNA in trypanosomid flagellates. **International review of cytology**, v. 99, p. 119–179, 1986.

SOKOL, D.; BENSON, G.; TOJEIRA, J. Tandem repeats over the edit distance. **Bioinformatics** , v. 23, n. 2, p. e30–5, 2007.

SOUZA, W. Basic Cell Biology of Trypanosoma cruzi. **Current Pharmaceutical Design**, v. 8, n. 4, p. 269–285, 2002.

SZKLARCZYK, R.; HERINGA, J. Tracking repeats using significance and transitivity. Bioinformatics , v. 20 Suppl 1, p. i311–7, 2004.

TAKEUCHI, O.; AKIRA, S. Pattern Recognition Receptors and Inflammation. **Cell**, v. 140, n. 6, p. 805–820, 2010.

Tandem Repeat Annotation Library (TRAL). **Documentation Page**. Disponível em: *<https://acg-team.github.io/tral/index.html>*. Acesso em: 21 mai. 2018

Tandem Repeat Finder Website - **Página de configuração** - Disponível em: *<https://tandem.bu.edu/trf/trf.unix.help.html>* Acesso em: 16 jan. 2018

TriTrypDB. **The Kinetoplastid Genomics Resource**. Disponível em: *<http://tritrypdb.org/tritrypdb/>.*Acesso em: 25 out. 2017

ULLU, E.; MATTHEWS, K. R.; TSCHUDI, C. Temporal order of RNA-processing reactions in trypanosomes: rapid trans splicing precedes polyadenylation of newly synthesized tubulin transcripts. **Molecular and cellular biology**, v. 13, n. 1, p. 720–725, 1993.

VAN DER PLOEG, L. H. Discontinuous transcription and splicing in trypanosomes. **Cell**, v. 47, n. 4, p. 479–480, 1986.

Venn diagram online tool. **Bioinformatics & Evolutionary Genomics, Ghent University, Bélgica.** Disponível em: *<http://bioinformatics.psb.ugent.be/webtools/Venn/>*. Acesso em: 09 mar. 2019

WHO. **World Health Organization website**.Disponível em: <https://www.who.int/chagas/en/> - Acesso em: 07 fev. 2019

ZINGALES, B. Trypanosoma cruzi genetic diversity: Something new for something known about Chagas disease manifestations, serodiagnosis and drug sensitivity. **Acta tropica**, v. 184, p. 38–52, 2018.

APÊNDICE 1 - PARÂMETROS DE ENTRADA DAS FERRAMENTAS DE ANOTAÇÃO DE *TANDEM REPEATS*

TRF v.4.09

<https://tandem.bu.edu/trf/trf.download.html>
Download: 21/11/2017

Comando completo: ./trf409.linux64 clean_gene_transcripts.fasta 2 7 7 80 10 36 2000 -h -d -ngs | tee output.out

Pesos dentro dos padrões sugeridos pela ferramenta.
Match: 2; **Mismatch:** 7, **Delta:** 7 (Mismatch e Delta são aplicados como valores negativos, conferem penalidades mais severas, portanto saídas são menos permissivas);

Probabilidades dentro dos limites sugeridos pela ferramenta.
PM (Probabilidade de match): 80 e **PI (Probabilidade de Indel):** 20.

Escore mínimo (Minscore): 36. Critério de corte, que aliado ao valor de match 2, requererá alinhamento de 18 NTs (ou 6 AAs, sem GAPs) para que TR seja mantido. Lembrar que TR deve ser sempre composto de no mínimo 2 repetições, portanto 9x2 AAs ou 6x3 AAs no exemplo acima.

Tamanho máximo: 2000. Critério máximo para tamanho de cada módulo de TR, mantive o máximo da ferramenta.

Parâmetros adicionais:
IMPORTANTE: Função do MATLAB requer que exatamente esses parâmetros sejam usados para formatar saída da ferramenta.
-h: Suprime a geração de arquivos HTML, não utilizados na análise;
-d: Gera arquivo de texto com o nome output.out;
-ngs: Gera arquivo de saída resumido.

T-REKS

<https://bioinfo.crbm.cnrs.fr/index.php?route=tools&tool=3>
Download: 21/11/2017

O programa foi baixado e executado localmente utilizando sua interface gráfica, pois para anotação de arquivos multifasta o website da ferramenta requer seu download. A versão baixada não permite parametrização. Conforme documentação do site, o tamanho mínimo do TR é 9 para TRs de AA único (*homorepeats*) e 14 para os demais. Documentação não informa qual a matriz de substituição utilizada pela ferramenta, porém versão online permite seleção de similaridade mínima, o que indica utilização de uma.

PROGERF
<http://200.131.37.155/ligp>
Download: 03/04/2018

Comando completo: perl progerf.pl -q clean_gene_proteins.fasta -o output.out –i 2 –y 7 –r 5 –g 5 –v 10 –d 30 –m p

O programa foi executado com os parâmetros abaixo:

Initial subsequence length = 2
Final subsequence length = 7
Degeneration(%) = 20
Overlap Allowed(%) = 10
Gaps between subsequences = 5
Minimal of repeats = 4

TRUST
<http://www.ibi.vu.nl/programs/trustwww/>
Download: 04/06/2018
Comando completo: java -cp . nl.vu.cs.align.SelfSimilarity -fasta clean_gene_proteins.fasta -matrix PAM120 -noseg -gapo 8 -gapx 2 -noforce | tee output/trs_pam120_noSEG.out

A ferramenta foi executada através da linha de comando utilizando a opção -noseg, que não atribui uma máscara de complexidade ao fasta.

APÊNDICE 2 - VALORES DE ACURÁCIA E SENSIBILIDADE OBTIDOS NOS TESTES DOS ALGORITMOS DE MACHINE LEARNING

As linhas em verde foram selecionadas para verificação dos Falso Negativos (FN) com a lista de TRs

#	Distribuição	Algoritmo	Atributos	Sensibilidade	Acurácia	Sensibilidade	Acurácia
1	Balanced	MLP	14.DIVCOV + SLIDESTATS + SLIDEFUZZY + NORMDIFF	0.77535	0.83812	0.79949	0.8288
2	Balanced	SVM	18.SLIDEFUZZY + COMBFEAT SLIDEFUZZY	0.77535	0.81723	0.84027	0.86151
3	Balanced	SVM	1.DIVCOV (5,6,7,8,9)	0.77336	0.80592	0.7876	0.80076
4	Balanced	MLP	18.SLIDEFUZZY + COMBFEAT SLIDEFUZZY	0.77137	0.80113	0.82073	0.81436
5	Balanced	SVM	16.DIVCOV + COMBFEAT DIVCOV	0.76541	0.80635	0.7876	0.80246
6	Balanced	MLP	7.SLIDESTATS (4)	0.76342	0.73368	0.7808	0.75701
7	Balanced	RUSBoost	15.DIVCOV + SLIDESTATS + SLIDEFUZZY + PHYCHEM30	0.76143	0.81854	0.82073	0.84197
8	Balanced	MLP	1.DIVCOV (5,6,7,8,9)	0.76143	0.80766	0.7842	0.80246
9	Balanced	SVM	10.SLIDEFUZZY (5)	0.75944	0.79025	0.84622	0.85981
10	Balanced	RUSBoost	7.SLIDESTATS (4)	0.75944	0.71497	0.78335	0.75701
11	Balanced	RUSBoost	11.DIVCOV + SLIDESTATS	0.75746	0.84552	0.79949	0.8407
12	Unbalanced	RUSBoost	7.SLIDESTATS (4)	0.75746	0.74064	0.77315	0.74841
13	Balanced	MLP	15.DIVCOV + SLIDESTATS + SLIDEFUZZY + PHYCHEM30	0.75547	0.84769	0.80374	0.84664
14	Balanced	SVM	7.SLIDESTATS (4)	0.75547	0.69191	0.83008	0.8254
15	Balanced	SVM	15.DIVCOV + SLIDESTATS + SLIDEFUZZY + PHYCHEM30	0.75149	0.84813	0.80969	0.85896
16	Balanced	MLP	12.DIVCOV + SLIDEFUZZY	0.75149	0.81462	0.7825	0.81181
17	Balanced	SVM	17.SLIDEFUZZY + COMBFEAT DIVCOV	0.75149	0.80983	0.82923	0.85854
18	Balanced	SVM	5.SLIDESTATS (3,4,5,6,10)	0.75149	0.8094	0.83432	0.86873
19	Balanced	MLP	11.DIVCOV + SLIDESTATS	0.7495	0.85727	0.79099	0.8407
20	Balanced	MLP	13.DIVCOV + SLIDESTATS+ SLIDEFUZZY	0.7495	0.84856	0.80204	0.84664
21	Balanced	MLP	5.SLIDESTATS (3,4,5,6,10)	0.7495	0.82332	0.77145	0.81903
22	Balanced	MLP	16.DIVCOV + COMBFEAT DIVCOV	0.7495	0.80418	0.79354	0.81521
23	Balanced	SVM	13.DIVCOV + SLIDESTATS+ SLIDEFUZZY	0.74751	0.84682	0.79949	0.84579
24	Balanced	MLP	9.SLIDEFUZZY (7)	0.74553	0.82898	0.78845	0.82116
25	Balanced	MLP	10.SLIDEFUZZY (5)	0.74553	0.82115	0.77825	0.80884
26	Balanced	SVM	9.SLIDEFUZZY (7)	0.74553	0.80809	0.81308	0.85004

#	Distribuição	Algoritmo	Atributos	Sensibilidade	Acurácia	Sensibilidade	Acurácia
27	Balanced	SVM	11.DIVCOV + SLIDESTATS	0.74155	0.85074	0.79184	0.84027
28	Balanced	RUSBoost	5.SLIDESTATS (3,4,5,6,10)	0.74155	0.83159	0.74936	0.81691
29	Balanced	MLP	19.SLIDEFUZZY + COMBFEAT DIVCOV+PHYCHEM30	0.74155	0.81419	0.77485	0.81011
30	Balanced	RUSBoost	6.SLIDESTATS (3)	0.73956	0.71453	0.76211	0.75234
31	Balanced	SVM	14.DIVCOV + SLIDESTATS + SLIDEFUZZY + NORMDIFF	0.73757	0.84813	0.79524	0.8424
32	Balanced	MLP	8.SLIDEFUZZY (7,5,6,3,8)	0.73757	0.82071	0.80629	0.82328
33	Balanced	MLP	17.SLIDEFUZZY + COMBFEAT DIVCOV	0.73559	0.83029	0.79609	0.83008
34	Balanced	SVM	8.SLIDEFUZZY (7,5,6,3,8)	0.7336	0.82289	0.80969	0.85302
35	Balanced	RUSBoost	8.SLIDEFUZZY (7,5,6,3,8)	0.72962	0.83594	0.77995	0.83432
36	Balanced	RUSBoost	17.SLIDEFUZZY + COMBFEAT DIVCOV	0.72962	0.83594	0.77995	0.83432
37	Balanced	RUSBoost	19.SLIDEFUZZY + COMBFEAT DIVCOV+PHYCHEM30	0.72962	0.83594	0.77995	0.83432
38	Balanced	SVM	6.SLIDESTATS (3)	0.72962	0.6906	0.79864	0.7876
39	Balanced	SVM	19.SLIDEFUZZY + COMBFEAT DIVCOV+PHYCHEM30	0.72763	0.80418	0.83008	0.86449
40	Unbalanced	RUSBoost	14.DIVCOV + SLIDESTATS + SLIDEFUZZY + NORMDIFF	0.72167	0.86466	0.7706	0.8799
41	Unbalanced	RUSBoost	13.DIVCOV + SLIDESTATS+ SLIDEFUZZY	0.71968	0.86423	0.76975	0.87896
42	Balanced	RUSBoost	13.DIVCOV + SLIDESTATS+ SLIDEFUZZY	0.71968	0.85335	0.76126	0.83602
43	Balanced	RUSBoost	14.DIVCOV + SLIDESTATS + SLIDEFUZZY + NORMDIFF	0.71968	0.85335	0.76126	0.83602
44	Balanced	RUSBoost	12.DIVCOV + SLIDEFUZZY	0.71769	0.85074	0.77995	0.83942
45	Unbalanced	RUSBoost	17.SLIDEFUZZY + COMBFEAT DIVCOV	0.71769	0.84943	0.7672	0.86554
46	Unbalanced	RUSBoost	8.SLIDEFUZZY (7,5,6,3,8)	0.71769	0.84813	0.7689	0.86498
47	Unbalanced	RUSBoost	18.SLIDEFUZZY + COMBFEAT SLIDEFUZZY	0.71769	0.84813	0.77145	0.86535
48	Balanced	RUSBoost	18.SLIDEFUZZY + COMBFEAT SLIDEFUZZY	0.71769	0.84726	0.76636	0.83347
49	Unbalanced	RUSBoost	11.DIVCOV + SLIDESTATS	0.71571	0.86466	0.76551	0.87728
50	Unbalanced	RUSBoost	15.DIVCOV + SLIDESTATS + SLIDEFUZZY + PHYCHEM30	0.71173	0.8651	0.7672	0.88288
51	Unbalanced	RUSBoost	19.SLIDEFUZZY + COMBFEAT DIVCOV+PHYCHEM30	0.71173	0.84682	0.7689	0.86703
52	Balanced	RUSBoost	10.SLIDEFUZZY (5)	0.71173	0.81593	0.78335	0.81946
53	Balanced	SVM	12.DIVCOV + SLIDEFUZZY	0.70974	0.84073	0.75956	0.81946
54	Unbalanced	RUSBoost	10.SLIDEFUZZY (5)	0.70775	0.85248	0.75191	0.86125
55	Balanced	RUSBoost	9.SLIDEFUZZY (7)	0.70577	0.83943	0.75276	0.8271
56	Unbalanced	RUSBoost	6.SLIDESTATS (3)	0.70577	0.7302	0.72642	0.74413
57	Unbalanced	RUSBoost	12.DIVCOV + SLIDEFUZZY	0.6998	0.85205	0.76296	0.87691
58	Balanced	MLP	6.SLIDESTATS (3)	0.68787	0.73064	0.68309	0.72387
59	Unbalanced	RUSBoost	5.SLIDESTATS (3,4,5,6,10)	0.67594	0.85466	0.70518	0.86367

#	Distribuição	Algoritmo	Atributos	Sensibilidade	Acurácia	Sensibilidade	Acurácia
60	Unbalanced	RUSBoost	9.SLIDEFUZZY (7)	0.67396	0.84769	0.74511	0.86833
61	Balanced	RUSBoost	1.DIVCOV (5,6,7,8,9)	0.65606	0.84943	0.70263	0.81223
62	Balanced	RUSBoost	4.PHYCHEM (30)	0.65408	0.60487	0.69499	0.66992
63	Unbalanced	RUSBoost	1.DIVCOV (5,6,7,8,9)	0.64414	0.84595	0.70858	0.86236
64	Unbalanced	RUSBoost	16.DIVCOV + COMBFEAT DIVCOV	0.63618	0.84639	0.69924	0.86908
65	Balanced	RUSBoost	16.DIVCOV + COMBFEAT DIVCOV	0.63419	0.84769	0.70603	0.81648
66	Balanced	MLP	4.PHYCHEM (30)	0.62624	0.65231	0.63042	0.65548
67	Unbalanced	MLP	15.DIVCOV + SLIDESTATS + SLIDEFUZZY + PHYCHEM30	0.59642	0.87032	0.66185	0.8909
68	Unbalanced	SVM	9.SLIDEFUZZY (7)	0.58648	0.86815	0.6695	0.91179
69	Unbalanced	SVM	15.DIVCOV + SLIDESTATS + SLIDEFUZZY + PHYCHEM30	0.57654	0.87772	0.6576	0.90395
70	Balanced	SVM	4.PHYCHEM (30)	0.57654	0.66406	0.84197	0.87086
71	Unbalanced	SVM	2.DIVCOV (6)	0.57455	0.85944	0.62617	0.87467
72	Unbalanced	RUSBoost	2.DIVCOV (6)	0.57455	0.85901	0.62872	0.8743
73	Balanced	RUSBoost	2.DIVCOV (6)	0.57455	0.85901	0.62872	0.7876
74	Balanced	MLP	2.DIVCOV (6)	0.57455	0.85857	0.62872	0.78632
75	Balanced	SVM	2.DIVCOV (6)	0.57455	0.85857	0.62872	0.78632
76	Unbalanced	MLP	2.DIVCOV (6)	0.56859	0.86031	0.62107	0.87449
77	Unbalanced	SVM	14.DIVCOV + SLIDESTATS + SLIDEFUZZY + NORMDIFF	0.56461	0.87511	0.64826	0.90433
78	Unbalanced	MLP	13.DIVCOV + SLIDESTATS+ SLIDEFUZZY	0.56262	0.87163	0.63042	0.89202
79	Unbalanced	MLP	8.SLIDEFUZZY (7,5,6,3,8)	0.56262	0.86771	0.63212	0.89202
80	Unbalanced	SVM	11.DIVCOV + SLIDESTATS	0.56064	0.8738	0.62362	0.89873
81	Unbalanced	SVM	18.SLIDEFUZZY + COMBFEAT SLIDEFUZZY	0.56064	0.86292	0.67799	0.91794
82	Unbalanced	MLP	5.SLIDESTATS (3,4,5,6,10)	0.56064	0.85074	0.59898	0.86665
83	Unbalanced	RUSBoost	4.PHYCHEM (30)	0.55865	0.70104	0.59898	0.72231
84	Unbalanced	SVM	13.DIVCOV + SLIDESTATS+ SLIDEFUZZY	0.55467	0.8725	0.61852	0.89071
85	Unbalanced	MLP	14.DIVCOV + SLIDESTATS + SLIDEFUZZY + NORMDIFF	0.55268	0.87032	0.61937	0.89109
86	Unbalanced	SVM	8.SLIDEFUZZY (7,5,6,3,8)	0.55268	0.8651	0.64316	0.90638
87	Unbalanced	SVM	10.SLIDEFUZZY (5)	0.54871	0.85074	0.69584	0.91999
88	Unbalanced	SVM	19.SLIDEFUZZY + COMBFEAT DIVCOV+PHYCHEM30	0.54672	0.85727	0.6593	0.91253
89	Unbalanced	MLP	11.DIVCOV + SLIDESTATS	0.54274	0.87163	0.60493	0.88941
90	Unbalanced	MLP	17.SLIDEFUZZY + COMBFEAT DIVCOV	0.54076	0.86292	0.61003	0.88941
91	Unbalanced	SVM	17.SLIDEFUZZY + COMBFEAT DIVCOV	0.53479	0.86336	0.63976	0.90545

#	Distribuição	Algoritmo	Atributos	Sensibilidade	Acurácia	Sensibilidade	Acurácia
92	Unbalanced	MLP	1.DIVCOV (5,6,7,8,9)	0.5169	0.8651	0.59303	0.884
93	Unbalanced	SVM	12.DIVCOV + SLIDEFUZZY	0.51491	0.86554	0.58794	0.88586
94	Unbalanced	MLP	18.SLIDEFUZZY + COMBFEAT SLIDEFUZZY	0.51491	0.86336	0.57179	0.88046
95	Unbalanced	SVM	5.SLIDESTATS (3,4,5,6,10)	0.51292	0.85161	0.6644	0.90507
96	Unbalanced	SVM	1.DIVCOV (5,6,7,8,9)	0.50298	0.86379	0.57859	0.88325
97	Unbalanced	SVM	16.DIVCOV + COMBFEAT DIVCOV	0.50298	0.86292	0.57859	0.88419
98	Unbalanced	MLP	10.SLIDEFUZZY (5)	0.49901	0.85814	0.565	0.87635
99	Unbalanced	MLP	16.DIVCOV + COMBFEAT DIVCOV	0.49304	0.86249	0.57944	0.8853
100	Unbalanced	MLP	12.DIVCOV + SLIDEFUZZY	0.48708	0.86031	0.58029	0.88568
101	Unbalanced	MLP	19.SLIDEFUZZY + COMBFEAT DIVCOV+PHYCHEM30	0.46918	0.85857	0.51317	0.873
102	Unbalanced	MLP	3.DIVCOV (7)	0.44533	0.86118	0.48513	0.87076
103	Unbalanced	SVM	3.DIVCOV (7)	0.44533	0.86118	0.48513	0.87076
104	Unbalanced	RUSBoost	3.DIVCOV (7)	0.44533	0.86118	0.48513	0.87076
105	Balanced	MLP	3.DIVCOV (7)	0.44533	0.86118	0.48513	0.73407
106	Balanced	SVM	3.DIVCOV (7)	0.44533	0.86118	0.48513	0.73407
107	Balanced	RUSBoost	3.DIVCOV (7)	0.44135	0.86031	0.48428	0.73534
108	Unbalanced	MLP	9.SLIDEFUZZY (7)	0.39761	0.84334	0.42396	0.8508
109	Unbalanced	MLP	6.SLIDESTATS (3)	0.30616	0.82376	0.35684	0.84017
110	Unbalanced	MLP	7.SLIDESTATS (4)	0.29821	0.81506	0.3305	0.82562
111	Unbalanced	SVM	7.SLIDESTATS (4)	0.27237	0.80374	0.49533	0.87766
112	Unbalanced	SVM	6.SLIDESTATS (3)	0.23658	0.80026	0.41206	0.86143
113	Unbalanced	MLP	4.PHYCHEM (30)	0.1829	0.802	0.22175	0.81742
114	Unbalanced	SVM	4.PHYCHEM (30)	0.16302	0.80026	0.42651	0.86945

ANEXO 1 - LISTA DE ANTÍGENOS RECOMBINANTES DE *T. CRUZI*

Hernández *et al.* (2010)realizaram um estudo onde os principais antígenos mapeados até então foram sintetizados em genes com até 9 regiões repetidas. Abaixo temos as sequências apresentadas neste estudo, que nos auxiliaram em algumas definições de filtragem de bons valores de probabilidade de regiões de epitopos de célula B:

MAP
Sequência:
PRHVDPDHFRSTTQDAYRPVDPSAYKRALPLEEEEDVG
PRHVDPDHFRSTTQDAYRPVDPSAYKRALPLEEEEDVG
PRHVDPDHFRSTTQDAYRPVDPSAYKRALPQEEEEDVG

JL8
Sequência:
AAEATKVAEAEKQR
AAEATKAVEAEKQR
AAEATKVAEAEKQK

CRA
Sequência:
KVAEAEKQKAAEAT
KVAEAEKQKAAEAT
KVAEAEKQKAAEAT

B13
Sequência:
PFGQAAAGDKPS
PFGQAAAGDKPS
PFGQAAAGDKPK

FRA
Sequência:
AFLDQKPEGVPLRELPLDDDSDFVAMEQERRQLLEKDPRRNAREIAAL
EESMNARAQELAREKKLADR
AFLDQKPEGVPLRELPLDDDSDFVAME
QERRQLLEKDPRRNAKEIAALEESMNARAQELAREKKLADR
AFLDQK

FRA2
Adaptado do FRA, região sublinhada contém um espaçador rico em prolina.
Sequência:
MEQERRQLLEKDPRRNAREIAALEESMNARAQELAREKKLADRAFPDSPNSMEQE
RRQLLEKDPRRNAREIAALEESMNARAQELAREKKLADRAFPNSPDMEQERRQLLE
KDPRRNAREIAALEESMNARAQELAREKKLADRAF

TcD
Sequência:
PKPAE
PKPAE
PKPAE
PKPAE
PKPAE
PKPAE
PKPAE
PKPAE
PKPAE

TcE
Sequência:
PAKAAA
PPAKAAA
PPAKAAA
PPAKAAA
PPAKAAA
PPAKAAAP

SAPA
Sequência:
PVDSSAHGTPST
PVDSSAHGTPST
PVDSSAHSTPST
PVDSSAHSTPST
PADSSAHSTPST

TcMyo
Sequência:
LAQREADNEKLAED
LAQREADNEKLAEE
LAQREADNEKLTED
LAQREADNEKLAED